Zero-Order One-Point Estimate with Distributed Stochastic Gradient-Tracking Technique

Anonymous authors
Paper under double-blind review

Abstract

In this work, we consider a distributed multi-agent stochastic optimization problem, where each agent holds a local objective function that is smooth and convex and that is subject to a stochastic process. The goal is for all agents to collaborate to find a common solution that optimizes the sum of these local functions. With the practical assumption that agents can only obtain noisy numerical function queries at precisely one point at a time, we extend the distributed stochastic gradient-tracking method to the bandit setting where we do not have an estimate of the gradient, and we introduce a zero-order (ZO) one-point estimate (1P-DSGT). We analyze the convergence of this novel technique for smooth and convex objectives using stochastic approximation tools, and we prove that it *converges almost surely to the optimum* despite the biasedness of our gradient estimate. We then study the convergence rate for when the objectives are additionally strongly convex. With constant step sizes, our method competes with its first-order (FO) counterparts by achieving a linear rate $O(\varrho^k)$ as a function of number of iterations k. To the best of our knowledge, this is the first work that proves this rate in the noisy estimation setting or with one-point estimators. With vanishing step sizes, we establish a rate of $O(\frac{1}{\sqrt{k}})$ after a sufficient number of iterations $k > K_2$. This is the optimal rate proven in the literature for centralized techniques utilizing one-point estimators. We then provide a regret bound of $O(\sqrt{k})$ with vanishing step sizes. We further illustrate the usefulness of the proposed technique using numerical experiments.

1 Introduction

Gradient-free optimization is an old topic in the research community; however, there has been an increased interest recently, especially in machine learning applications, where optimization problems are typically solved with gradient descent algorithms. Successful applications of gradient-free methods in machine learning include competing with an adversary in bandit problems (Flaxman et al., 2004; Agarwal et al., 2010), generating adversarial attacks for deep learning models (Chen et al., 2019; Liu et al., 2019) and reinforcement learning (Vemula et al., 2019). Gradient-free optimization aims to solve optimization problems with only functional ZO information rather than FO gradient information. These techniques are essential in settings where explicit gradient computation may be impractical, expensive, or impossible. Instances of such settings include high data dimensionality, time or resource straining function differentiation, or the cost function not having a closed-form. ZO information-based methods include direct search methods (Golovin et al., 2019), 1-point methods (Flaxman et al., 2004; Bach & Perchet, 2016; Vemula et al., 2019; Li & Assaad, 2021) where a function $f(\cdot, S) : \mathbb{R}^d \to \mathbb{R}$ is evaluated at a single point with some randomization to estimate the gradient as such

$$g^{(1)}_{\gamma,z}(x, S) = \frac{d}{\gamma} f(x + \gamma z, S) z, \tag{1}$$

with x the optimization variable, $\gamma > 0$ a small value, and z a random vector following a symmetrical distribution. ZO also includes 2- or more point methods (Duchi et al., 2015; Nesterov & Spokoiny, 2017; Gorbunov et al., 2018; Bach & Perchet, 2016; Hajinezhad et al., 2019; Kumar Sahu et al., 2018; Agarwal et al., 2010; Chen et al., 2019; Liu et al., 2019; Vemula et al., 2019), where functional difference at various

points is employed for estimation, generally having the respective structures

$$g^{(2)}_{\gamma,z}(x,S) = d\frac{f(x+\gamma z,S)-f(x-\gamma z,S)}{2\gamma}z \tag{2}$$

$$\text{and } g^{(2d)}_{\gamma}(x,S) = \sum_{j=1}^{d}\frac{f(x+\gamma e_j,S)-f(x-\gamma e_j,S)}{2\gamma}e_j \tag{3}$$

where $\{e_j\}_{j=1,\dots,d}$ is the canonical basis, and other methods such as sign information of gradient estimates (Liu et al., 2019).

Another area of great interest is distributed multi-agent optimization, where agents try to cooperatively solve a problem with information exchange only limited to immediate neighbors in the network. Distributed computing and data storing are particularly essential in fields such as vehicular communications and coordination, data processing and distributed control in sensor networks (Shi & Eryilmaz, 2020), big-data analytics (Daneshmand et al., 2015), and federated learning (McMahan et al., 2017). More specifically, one direction of research integrates (sub)gradient-based methods with a consensus/averaging strategy; the local agent incorporates one or multiple consensus steps alongside evaluating the local gradient during optimization. Hence, these algorithms can tackle a fundamental challenge: overcoming differences between agents' local data distributions.

1.1 Problem Description

Consider a set of agents $\mathcal{N} = \{1, 2, \dots, n\}$ connected by a communication network. Each agent i is associated with a local objective function $f_i(\cdot, S) : \mathbb{R}^d \to \mathbb{R}$. The global goal of the agents is to collaboratively locate the decision variable $x \in \mathbb{R}^d$ that solves the stochastic optimization problem:

$$min_{x\in\mathbb{R}^d}\mathcal{F}(x) = \frac{1}{n}\sum_{i=1}^{n}F_i(x) \tag{4}$$

where

$$F_i(x) = \mathbb{E}_S f_i(x,S),$$

with $S \in \mathcal{S}$ denoting an i.i.d. ergodic stochastic process describing uncertainties in the communication system.

We assume that at each time step, agent i can only query the function values of f_i at exactly one point, and can only communicate with its neighbors. Further, we assume that the function queries are noisy $\tilde{f}_i = f_i + \zeta_i$ with ζ_i some additive noise. Agent i must then employ this query to estimate the gradient of the form $g_i(x, S_i)$.

One efficient algorithm with a straightforward averaging scheme to solve this problem is the gradient-tracking (GT) technique, which has proven to achieve rates competing with its centralized counterparts. For example, the acquired error bound under a distributed stochastic variant was found to decrease with the network size n (Pu & Nedić, 2018). In most work, this technique proved to converge linearly to the optimal solution with constant step size (Qu & Li, 2018; Nedić et al., 2017; Pu, 2020), which is also a unique attribute among other distributed stochastic gradient algorithms. It has been extended to time-varying (undirected or directed) graphs (Nedić et al., 2017), a gossip-like method which is efficient in communication (Pu & Nedić, 2018), and nonconvex settings (Tang et al., 2021; Lorenzo & Scutari, 2016; Jiang et al., 2022; Lu et al., 2019). All references mentioned above consider the case where an accurate gradient computation or a non-biased gradient estimation with bounded variance (BV) is available, the variance being the mean squared approximation error of the gradient estimate with respect to the true gradient.

1.2 Function Classes and Gradient Estimate Assumptions

Consider the following five classes of functions:

- The convex class $\mathcal{C}_{cvx}$ containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are convex.
- The strongly convex class $\mathcal{C}_{sc}$ containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are continuously differentiable and admit a constant λ_f such that

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \| \geq \lambda_f \|x - y\|^2, \ \ \forall x, y \in \mathbb{R}^d.$$

- The Lipschitz continuous class $\mathcal{C}_{lip}$ containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ that admit a constant L_f such that

$$|f(x) - f(y)| \leq L_f \|x - y\|, \ \ \forall x, y \in \mathbb{R}^d.$$

- The smooth class $\mathcal{C}_{smo}$ containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are continuously differentiable and admit a constant G_f such that

$$\|\nabla f(x) - \nabla f(y)\| \leq G_f \|x - y\|, \ \ \forall x, y \in \mathbb{R}^d.$$

- The gradient dominated class $\mathcal{C}_{gd}$ containing all functions $f : \mathbb{R}^d \to \mathbb{R}$ that are differentiable, have a global minimizer x^*, and admit a constant ν_f such that

$$2\nu_f(f(x) - f(x^*)) \leq \|\nabla f(x)\|^2, \ \ \forall x \in \mathbb{R}^d.$$

  This gradient domination property can be viewed as a nonconvex analogy of strong convexity.

In addition, consider the following assumptions on the gradient estimate:

- A gradient estimate g is said to be unbiased w.r.t. the true gradient ∇f if for all $x \in \mathbb{R}^d$ and independent $S \in \mathcal{S}$, it satisfies the following equality

$$\mathbb{E}_S[g(x, S)|x] = \nabla f(x).$$

- Otherwise, it is said to be biased and satisfies

$$\mathbb{E}_S[g(x, S)|x] = \nabla f(x) + b(x),$$

  with $b(x)$ some bias term.
- A gradient estimate g is said to have bounded variance when for all $x \in \mathbb{R}^d$ and independent $S \in \mathcal{S}$,

$$\mathbb{E}_S[\|g(x, S) - \nabla f(x)\|^2|x] \leq \sigma \ \text{ for some } \sigma > 0.$$

- Otherwise, when this bound is unknown or does not exist, it is said to have unbounded variance.

In general, FO stochastic gradient estimates are unbiased and have bounded variance. ZO estimates, on the other hand, are biased. While multi-point ZO estimates have bounded or even vanishing variance, one-point estimates have unbounded variance Liu et al. (2020).

1.3 Related Work

The optimal convergence rate for solving problem (4), assuming the objective function $\mathcal{F}$ is strongly convex with Lipschitz continuous gradients, has been established as $O(\frac{1}{k})$ under a diminishing step size with full gradient information Pu & Nedić (2018); Nemirovski et al. (2009). However, when employing a constant step size $\alpha > 0$ that is sufficiently small, the iterates produced by a stochastic gradient method converge exponentially fast (in expectation) to an $O(\alpha)$-neighborhood of the optimal solution (Pu & Nedić, 2018); this is known as the linear rate $O(\varrho^k)$. To solve problem (4), all Qu & Li (2018); Lorenzo & Scutari (2016); Nedić et al. (2017); Shi et al. (2015); Li et al. (2022); Jiang et al. (2022) present a distributed gradient-tracking method that employs local auxiliary variables to track the average of all agents' gradients,

	GR. ES.	OP	FUNCTION CLASS	CONS-ENSUS	STEP SIZE	REGRET BOUND	CONVERGENCE RATE
ZO	One-point	Cent.	$\mathcal{C}_{cvx} \bigcap \mathcal{C}_{lip}$	-	f.	$O(k^{\frac{3}{4}})$	$O(\frac{1}{\sqrt[4]{k}})$ Flaxman et al. (2004)
		Dist.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{lip} \bigcap \mathcal{C}_{smo}$	GT	v.	$O(\sqrt{k})$	$O(\frac{1}{\sqrt{k}})$ 1P-DSGT
		Dist.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{lip} \bigcap \mathcal{C}_{smo}$	GT	f.	-	$O(\varrho^k)$ 1P-DSGT
	Two-point	Cent.	$\mathcal{C}_{cvx} \bigcap \mathcal{C}_{lip}$	-	v.	$O(\sqrt{k})$	$O(\frac{1}{\sqrt{k}})$ Agarwal et al. (2010)
		Cent.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{lip}$	-	v.	$O(\log k)$	$O(\frac{\log k}{k})$Agarwal et al. (2010)
		Dist.	$\mathcal{C}_{lip} \bigcap \mathcal{C}_{smo}$	None	v.	-	$O(\frac{1}{\sqrt{k}} \log k)$ Tang et al. (2021)
		Dist.	$\mathcal{C}_{smo} \bigcap \mathcal{C}_{gd}$	None	v.	-	$O(\frac{1}{k})$ Tang et al. (2021)
	(d+1)-point	Cent.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{lip} \bigcap \mathcal{C}_{smo}$	-	v.	$O(\log k)$	$O(\frac{\log k}{k})$ Agarwal et al. (2010)
	2d-point	Dist.	$\mathcal{C}_{smo}$	GT	f.	-	$O(\frac{1}{k})$ Tang et al. (2021)
		Dist.	$\mathcal{C}_{smo} \bigcap \mathcal{C}_{gd}$	GT	f.	-	$O(\varrho^k)$ Tang et al. (2021)
	Kernel -based	Cent.	$\mathcal{C}_{cvx} \bigcap \mathcal{C}_{lip}$	-	v.	$O(\sqrt{k})$	$O(\frac{1}{\sqrt{k}})$ Bubeck et al. (2021)
FO	Unbiased /BV	Dist.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{smo}$	GT	f.	-	$O(\varrho^k)$ Pu & Nedić (2018); Xin et al. (2019); Pu (2020)
		Dist.	$\mathcal{C}_{sc} \bigcap \mathcal{C}_{smo}$	GT	v.	-	$O(\frac{1}{k})$ Pu & Nedić (2018)
		Dist.	$\mathcal{C}_{smo}$	GT	v.	-	$O(\frac{1}{\sqrt{k}})$ Lu et al. (2019)

Table 1: Convergence rates for various algorithms related to our work, classified according to the nature of the gradient estimate (gr. es.), whether the optimization problem (OP) is centralized or distributed, the assumptions on the objective function, whether consensus is aimed at or not, whether the step size is fixed (f.) or varying (v.), and the achieved regret bound and convergence rate

considering the availability of accurate gradient information. In both Li et al. (2022); Jiang et al. (2022), each local objective function is an average of finite instantaneous functions. Thus, they incorporate the gradient-tracking algorithm with stochastic averaging gradient technology (Li et al., 2022) (smooth convex optimization) or with variance reduction techniques (Jiang et al., 2022) (smooth nonconvex optimization). At each iteration, they randomly select only one gradient of an instantaneous function to approximate the local batch gradient. In Li et al. (2022), this is an unbiased estimate of the local gradient, whereas, in Jiang et al. (2022), it is biased. Nevertheless, both references assume access to an exact gradient oracle.

All Pu & Nedić (2018); Xin et al. (2019); Pu (2020); Lu et al. (2019) assume access to local stochastic FO oracles where the gradient is unbiased and with a bounded variance. In the first three, they additionally assume smooth and strongly-convex local objectives, and they all accomplish a linear convergence rate under a constant step size. Pu & Nedić (2018) propose a distributed stochastic gradient-tracking method (DSGT) and a gossip-like stochastic gradient-tracking method (GSGT) where at each round, each agent wakes up with a certain probability. Further, in Pu & Nedić (2018), when the step-size is diminishing, the convergence rate is that of $O(\frac{1}{k})$. Xin et al. (2019) employ a gradient-tracking algorithm for agents communicating over a strongly-connected graph. Pu (2020) introduces a robust gradient-tracking method (R-Push-Pull) in the context of noisy information exchange between agents and with a directed network topology. Lu et al. (2019) propose a gradient-tracking based nonconvex stochastic decentralized (GNSD) algorithm for

nonconvex optimization problems in machine learning, and they fulfill a convergence rate of $O(\frac{1}{\sqrt{k}})$ under constant step size.

On the other hand, ZO methods are known to have worse convergence rates than their FO counterparts under the same conditions. For example, under a convex centralized setting, Flaxman et al. (2004) prove a regret bound of $O(k^{\frac{3}{4}})$ (or equivalently a rate of $O(\frac{1}{\sqrt[4]{k}})$) with a one-point estimator for Lipschitz continuous functions. In the work of Agarwal et al. (2010), when the number of points is two, they prove regret bounds of $\tilde{O}(\sqrt{k})$ with high probability and of $O(\log(k))$ in expectation for strongly convex loss functions. When the number is $d+1$ point, they prove regret bounds of $O(\sqrt{k})$ and of $O(\log(k))$ with strong convexity. The reason why the performance improves with the addition of number of points in the estimate, is that their variance can be bounded, unlike one-point estimates whose variance cannot be bounded (Liu et al., 2020). However, when the function queries are subjected to noise, multi-point estimates start behaving like one-point ones. In noisy function queries (centralized) scenario, it has been proven that gradient-free methods cannot achieve a better convergence rate than $\Omega(\frac{1}{\sqrt{k}})$ which is the lower bound derived by Duchi et al. (2015); Jamieson et al. (2012); Shamir (2013) for strongly convex and smooth objective functions. In the work of Bubeck et al. (2021), a kernelized loss estimator is proposed where a generalization of Bernoulli convolutions is adopted, and an annealing schedule for exponential weights is used to control the estimator's variance in a focus region for dimensions higher than 1. Their method achieves a regret bound of $O(\sqrt{k})$.

In distributed settings, Tang et al. (2021) develop two algorithms for a noise-free nonconvex multi-agent optimization problem aiming at consensus. One of them is gradient-tracking based on a 2d-point estimator of the gradient with vanishing variance that achieves a rate of $O(\frac{1}{k})$ with smoothness assumptions and a linear rate for an extra ν-gradient dominated objective assumption and for fixed step sizes. The other is based on a 2-point estimator without global gradient tracking and achieves a rate of $O(\frac{1}{\sqrt{k}}\log k)$ under Lipschitz continuity and smoothness conditions and $O(\frac{1}{k})$ under an extra gradient dominated function structure.

We summarize all the mentioned convergence rates from the literature in Table 1.

1.4 Contributions

While the gradient tracking method has been extended to the ZO case (Tang et al., 2021), the approach followed by Tang et al. (2021) relies on a multi-point gradient estimator. It also assumes a static objective function devoid of any stochasticity or noise in the system. However, real-world scenarios often involve various sources of stochasticity and noise, such as differing data distributions among devices, perturbations in electronic components, quantization errors, data compression losses, and fluctuations in communication channels over time. Consequently, static objective functions become inadequate for realistic modeling. Moreover, the multi-point estimation technique assumes the ability to observe multiple instances of the objective function under identical system conditions, i.e., many function queries are done for the same realization of S in (2) and (3). However, this assumption needs to be revised in applications such as mobile edge computing (Mao et al., 2017; Chen et al., 2021; Zhou et al., 2022) where computational tasks from mobile users are offloaded to servers within the cellular network. Thus, queries requested from the servers by the users are subject to the wireless environment and are corrupted by noise not necessarily additive. Other applications include sensor selection for an accurate parameter estimation (Liu et al., 2018) where the observation of each sensor is continuously changing. Thus, in such scenarios, one-point estimates offer a vital alternative to solving online optimization/learning problems. Yet, one-point estimators are not generally used because of their slow convergence rate. The main reason is due to their unbounded variance. To avoid this unbounded variance, in this work, we don't use the estimate given in (1), we extend the one point approach in Li & Assaad (2021)'s work where the action of the agent is a scalar and different agents have different variables, to our consensual problem with vector variables. The difference is that in our gradient estimate, we don't divide by γ. This brings additional challenges in proving that our algorithm converges and a consensus can be achieved by all agents. And even with bounded variance, there's still a difficulty achieving good convergence rates when combining two-point estimates with the gradient tracking method due to the constant upper bound of the variance (Tang et al., 2021). Here, despite this constant bound, we were able to go beyond two-point estimates and achieve a linear rate. Moreover, while it requires $2d$ points to achieve a linear rate

in Tang et al. (2021)'s work, which is twice the dimension of the gradient itself, here we only need one scalar point or query. This is much more computationally efficient.

We summarize our contribution in the following points,

- We consider smooth and convex local objectives, and we extend the gradient-tracking algorithm to the case where we do not have an estimation of the gradient in the noisy setting.
- Under the realistic assumption that the agent only has access to a single noisy function value at each time without necessarily knowing the form of this function, we propose a one-point estimator in a stochastic framework.
- Naturally, one-point estimators are biased with respect to the true gradient and suffer from high variance (Liu et al., 2020) hence they do not match the assumptions for convergence presented in Tang et al. (2021); Pu & Nedić (2018); Xin et al. (2019); Pu (2020); Lu et al. (2019). However, in this work, we analyze and indeed prove the algorithm's convergence almost surely with a biased estimate. We also consider that a stochastic process influences the objective function from one iteration to the other, which is not the case in the aforementioned references.
- We then study the convergence rate for smooth and strongly convex objectives and we demonstrate that with fixed step sizes, the algorithm achieves a linear convergence rate $O(\varrho^k)$, marking the first instance where this rate is attained in ZO optimization with one-point/two-point estimates and in a noisy query setting, to the best of our knowledge. This linear rate competes with FO methods and even centralized algorithms in terms of convergence speed (Pu & Nedić, 2018).
- When the step-sizes are vanishing, we prove that a rate of $O(\frac{1}{\sqrt{k}})$ after a sufficient number of iterations $k > K_2$ is attainable. This rate satisfies the lower bounds achieved by its centralized counterparts in the same derivative-free setting (Duchi et al., 2015; Jamieson et al., 2012; Shamir, 2013).
- We then show that a regret bound of $O(\sqrt{k})$ is achieved for this algorithm.
- Finally, we support our theoretical claims by providing numerical evidence and comparing the algorithm's performance to its FO counterpart.

The rest of this paper is divided as follow. In subsection 1.5, we present the mathematical notation followed in this paper. In subsection 1.6, we present the main assumptions of our optimization problem. We then describe our gradient estimate followed by the proposed algorithm in section 2.1. We then prove the almost sure convergence of our algorithm in subsection 2.2 and study its rate in subsection 2.3 with varying step sizes. In subsection 2.4, we find its regret bound. And in subsection 2.5, we consider the case of fixed step sizes, study the convergence of our algorithm and its rate. Finally, in section 3, we provide numerical evidence and conclude the paper in section 4.

1.5 Notation

In all that follows, vectors are column-shaped unless defined otherwise and $\mathbf{1}$ denotes the vector of all entries equal to 1. For two vectors a, b of the same dimension, $\langle a, b\rangle$ is the inner product. For two matrices A, $B \in \mathbb{R}^{n\times d}$, we define

$$\langle A, B\rangle = \sum_{i=1}^{n} \langle A_i, B_i\rangle$$

where A_i (respectively, B_i) represents the i-th row of A (respectively, B). $\|.\|$ denotes the 2-norm for vectors and the Frobenius norm for matrices.

We assume that each agent i maintains a local copy $x_i \in \mathbb{R}^d$ of the decision variable and another auxiliary variable $y_i \in \mathbb{R}^d$ and each agent's local function is subject to the stochastic variable $S_i \in \mathbb{R}^m$. At iteration

k, the respective values are denoted as $x_{i,k}$, $y_{i,k}$, and $S_{i,k}$. Bold notations denote the concatenated version of the variables, i.e.,

$$\mathbf{x} := [x_1, x_2, \ldots, x_n]^T,\ \mathbf{y} := [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^{n\times d},\ \text{and}\ \mathbf{S} := [S_1, S_2, \ldots, S_n]^T \in \mathbb{R}^{n\times m}.$$

We then define the means of the previous two variables as $\bar{x} := \frac{1}{n}\mathbf{1}^T\mathbf{x}$ and $\bar{y} := \frac{1}{n}\mathbf{1}^T\mathbf{y} \in \mathbb{R}^{1\times d}$.

We define the gradient of F_i at the local variable $\nabla F_i(x_i) \in \mathbb{R}^d$ and its Hessian matrix $\nabla^2 F_i(x_i) \in \mathbb{R}^{d\times d}$ and we let

$$\nabla F(\mathbf{x}) := [\nabla F_1(x_1), \nabla F_2(x_2), \ldots, \nabla F_n(x_n)]^T \in \mathbb{R}^{n\times d}$$

and

$$\mathbf{g} := g(\mathbf{x}, \mathbf{S}) := [g_1(x_1, S_1), g_2(x_2, S_2), \ldots, g_n(x_n, S_n)]^T \in \mathbb{R}^{n\times d}.$$

We define its mean $\bar{g} := \frac{1}{n}\mathbf{1}^T\mathbf{g} \in \mathbb{R}^{1\times d}$ and we denote each agent's gradient estimate at time k by $g_{i,k} = g_i(x_{i,k}, S_{i,k})$.

1.6 Basic Assumptions

In this subsection, we introduce the fundamental assumptions that ensure the performance of the 1P-DSGT algorithm.

Assumption 1.1. *(on the graph) The topology of the network is represented by the graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where the edges in $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ represent communication links. The graph $\mathcal{G}$ is undirected, i.e., $(i, j) \in \mathcal{E}$ iff $(j, i) \in \mathcal{E}$, and connected (there exists a path of links between any two agents).*

$W = [w_{ij}] \in \mathbb{R}^{n\times n}$ denotes the agents' coupling matrix, where agents i and j are connected iff $w_{ij} = w_{ji} > 0$ ($w_{ij} = w_{ji} = 0$ otherwise). W is a nonnegative matrix and doubly stochastic, i.e., $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T W = \mathbf{1}^T$. All diagonal elements w_{ii} are strictly positive.

Assumption 1.2. *(on the objective function) We assume the existence and the continuity of both $\nabla F_i(x)$ and $\nabla^2 F_i(x)$. Let $x^* \in \mathbb{R}^d$ denote the solution of the problem (4), then $\nabla F_i(x^*) = 0$ and $\det(\nabla^2 F_i(x^*)) > 0$, $\forall i \in \mathcal{N}$. To insure the existence of x^*, we let the objective function be strictly convex, i.e.,*

$$\langle x - x^*, \nabla \mathcal{F}(x)\rangle \geq 0, \forall x \in \mathbb{R}^d. \tag{5}$$

We further assume the boundedness of the local Hessian where there exists a constant $\alpha_1 \in \mathbb{R}^+$ such that

$$\|\nabla^2 F_i(x)\|_2 \leq \alpha_1,\ \forall i \in \mathcal{N},$$

where here it suffices to use the Euclidean norm for matrices (keeping in mind for a matrix A, $\|A\|_2 \leq \|A\|_F$).

Assumption 1.3. *(on the local functions) All local functions $x \longmapsto f_i(x, S)$ are Lipschitz continuous with Lipschitz constant L_S,*

$$\|f_i(x, S) - f_i(x', S)\| \leq L_S\|x - x'\|,\ \forall i \in \mathcal{N}.$$

In addition, we assume $\mathbb{E}_S f_i(x, S) < \infty$, $\forall i \in \mathcal{N}$, to guarantee the boundedness of the objective $\mathcal{F}(x)$.

Assumption 1.4. *(on the additive noise) $\zeta_{i,k}$ is a zero-mean uncorrelated noise with bounded variance, where $E(\zeta_{i,k}) = 0$, $E(\zeta_{i,k}^2) = \alpha_4 < \infty$, $\forall i \in \mathcal{N}$, and $E(\zeta_{i,k}\zeta_{j,k}) = 0$ if $i \neq j$.*

Lemma 1.5. *(Qu & Li, 2018) Let ρ_w be the spectral norm $W - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. When Assumption 1.1 is satisfied, we have the following inequality*

$$\|W\omega - \mathbf{1}\bar{\omega}\| \leq \rho_w\|\omega - \mathbf{1}\bar{\omega}\|,\ \forall \omega \in \mathbb{R}^{n\times d}\ \text{and}\ \bar{\omega} = \frac{1}{n}\mathbf{1}^T\omega,$$

and $\rho_w < 1$.

Lemma 1.6. *Define $h(\mathbf{x}) := \frac{1}{n}\mathbf{1}^T\nabla F(\mathbf{x}) \in \mathbb{R}^{1\times d}$. Due to the boundedness of the second derivative in Assumption 1.2, the objective function is thus L-smooth and we have*

$$\|\nabla \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\| \leq \frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|.$$

2 Distributed Stochastic Gradient-Tracking Method

We propose to employ a zero-order one-point estimate of the gradient subject to the stochastic process S and an additive noise ζ while a stochastic perturbation and a step size are introduced, and we assume that each agent can perform this estimation at each iteration. To elaborate, let $g_{i,k}$ denote the aforementioned gradient estimate for agent i at time k, then we define it as

$$\begin{aligned} g_{i,k} &= \Phi_{i,k}\tilde{f}_i(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k}) \\ &= \Phi_{i,k}(f_i(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k}) + \zeta_{i,k}), \end{aligned} \tag{6}$$

where $\gamma_k > 0$ is a vanishing step size and $\Phi_{i,k} \in \mathbb{R}^d$ is a perturbation randomly and independently generated by each agent i. $g_{i,k}$ is in fact a biased estimation of the gradient $\nabla F_i(x_{i,k})$ and the algorithm can converge under the condition that all parameters are properly chosen. For clarification on the form of this bias and more on the properties of this estimate, refer to Appendix A.

2.1 The 1P-DSGT Algorithm

The following distributed stochastic gradient-tracking method is considered in this part making use of the gradient estimate presented in (6).

Every agent i initializes its variables with an arbitrary value $x_{i,0}$ and $y_{i,0} = g_{i,0}$. Then, at each time $k \in \mathbb{N}$, agent i updates its variables independently according to the following 3 steps:

$$\begin{aligned} &x_{i,k+1} = \sum_{j=1}^{n} w_{ij}(x_{j,k} - \alpha_k y_{j,k}) \\ &\text{perform the action: } x_{i,k+1} + \gamma_{k+1}\Phi_{i,k+1} \\ &y_{i,k+1} = \sum_{j=1}^{n} w_{ij} y_{j,k} + g_{i,k+1} - g_{i,k} \end{aligned} \tag{7}$$

where $\alpha_k > 0$ is a vanishing step size. Algorithm (7) can then be written in the following compact matrix form for clarity of analysis:

$$\begin{aligned} &\mathbf{x}_{k+1} = W(\mathbf{x}_k - \alpha_k \mathbf{y}_k) \\ &\text{perform the action: } \mathbf{x}_{k+1} + \gamma_{k+1}\mathbf{\Phi}_{k+1} \\ &\mathbf{y}_{k+1} = W\mathbf{y}_k + \mathbf{g}_{k+1} - \mathbf{g}_k \end{aligned} \tag{8}$$

where $\mathbf{\Phi}_k \in \mathbb{R}^{n\times d}$ is defined as $\mathbf{\Phi}_k = [\Phi_{1,k}, \Phi_{2,k}, \ldots, \Phi_{n,k}]^T$.

As is evident from the update of the variables, the exchange between agents is limited to neighboring nodes, and it encompasses the decision variable $\mathbf{x}_{k+1}$ and the auxiliary variable $\mathbf{y}_{k+1}$.

By construction of Algorithm (8), we note that the mean of the auxiliary variable $\mathbf{y}_k$ is equal to that of the gradient estimate $\mathbf{g}_k$ at every iteration k since $\mathbf{y}_0 = \mathbf{g}_0$, and by recursion, we obtain $\bar{y}_k = \frac{1}{n}\mathbf{1}^T\mathbf{g}_k = \bar{g}_k$.

The following assumption is only taken into account when we study the algorithm's behavior with varying step sizes, otherwise it is dropped.

Assumption 2.1. *(on the step-sizes) Both α_k and γ_k vanish to 0 as $k \to \infty$, and satisty the the following sums*

$$\sum_{k=1}^{\infty} \alpha_k\gamma_k = \infty, \text{ and } \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Assumption 2.2. *(on the random perturbation) Let $\Phi_{i,k} = (\phi_{i,k}^1, \phi_{i,k}^2, \ldots, \phi_{i,k}^d)^T$.*

Each agent i chooses its $\Phi_{i,k}$ vector independently from other agents $j \neq i$. In addition, the elements of $\Phi_{i,k}$ are assumed i.i.d with $\mathbb{E}(\phi_{i,k}^{d_1}\phi_{i,k}^{d_2}) = 0$ for $d_1 \neq d_2$ and there exists $\alpha_2 > 0$ such that $\mathbb{E}(\phi_{i,k}^{d_j})^2 = \alpha_2$, $\forall d_j$, $\forall i$. We further assume that there exists a constant $\alpha_3 > 0$ where $\|\Phi_{i,k}\| \leq \alpha_3$, $\forall i$.

Example 2.3. *One example is to take $\alpha_k = \alpha_0(k+1)^{-\upsilon_1}$ and $\gamma_k = \gamma_0(k+1)^{-\upsilon_2}$ with the constants α_0, γ_0, υ_1, $\upsilon_2 \in \mathbb{R}^+$. As $\sum_{k=1}^{\infty} \alpha_k \gamma_k$ diverges for $\upsilon_1 + \upsilon_2 \leq 1$ and $\sum_{k=1}^{\infty} \alpha_k^2$ converges for $\upsilon_1 > 0.5$, we can find pairs of υ_1 and υ_2 so that Assumption 2.1 is satisfied.*

To achieve the conditions in Assumption 2.2, we can choose the probability distribution of $\phi_{i,k}^{d_j}$ to be the symmetrical Bernoulli distribution where $\phi_{i,k}^{d_j} \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}$ with $\mathbb{P}(\phi_{i,k}^{d_j} = -\frac{1}{\sqrt{d}}) = \mathbb{P}(\phi_{i,k}^{d_j} = \frac{1}{\sqrt{d}}) = 0.5$, $\forall d_j$, $\forall i$.

2.2 Convergence Results

In this part, we analyze the asymptotic behavior of Algorithm (8). We start the analysis by defining $\mathcal{H}_k$ as the history sequence $\{x_0, y_0, S_0, \ldots, x_{k-1}, y_{k-1}, S_{k-1}, x_k\}$ and denoting by $\mathbb{E}[.|\mathcal{H}_k]$ as the conditional expectation given $\mathcal{H}_k$.

We define $\tilde{g}_k$ to be the expected value of $\bar{g}_k$ with respect to all the stochastic terms S, Φ, ζ given $\mathcal{H}_k$, i.e.,

$$\tilde{g}_k = \mathbb{E}_{S,\Phi,\zeta}[\bar{g}_k|\mathcal{H}_k]$$

In what follows, we use $\tilde{g}_k = \mathbb{E}[\bar{g}_k|\mathcal{H}_k]$ for shorthand notation.

We define the error e_k to be the difference between the value of a single realization of $\bar{g}_k$ and its conditional expectation $\tilde{g}_k$, i.e.,

$$e_k = \bar{g}_k - \tilde{g}_k,$$

where e_k can be seen as a stochastic noise. The following lemma describing the vanishing of the stochastic noise is essential for our main result.

Lemma 2.4. *If all Assumptions 1.3, 1.4, 2.1, and 2.2 hold and $\|\bar{x}_k\| < \infty$ almost surely, then for any constant $\nu > 0$, we have*

$$\mathbb{P}(\lim_{K\to\infty} \sup_{K' \geq K} \| \sum_{k=K}^{K'} \alpha_k e_k \| \geq \nu) = 0, \ \forall \nu > 0.$$

Proof: See Appendix B.

For any integer $k \geq 0$, we define the divergence, or the error between the average action taken by the agents $\bar{x}_k$ and the optimal solution x^* as

$$d_k = \|\bar{x}_k - x^*\|^2. \tag{9}$$

The following theorem describes the main convergence result.

Theorem 2.5. *If all Assumptions 1.1-1.4 and 2.1-2.2 hold and $\|\mathbf{x}_k\| < \infty$ almost surely, then as $k \to \infty$, $d_k \to 0$, $\bar{x}_k \to x^*$, and $x_{i,k} \to \bar{x}_k$ for all $i \in \mathcal{N}$ almost surely by applying the Algorithm.*

Proof: See Appendix C.

2.3 Convergence Rate with Vanishing Step Sizes

This part deals with how fast the expected divergence vanishes to find the proposed algorithm's expected convergence rate. To do so, we define the expected divergence to be $D_k = \mathbb{E}[\|\bar{x}_k - x^*\|^2]$. The goal is to bound this divergence from above by sequences whose convergence rate is known. The analysis is highly associated with the parameters α_k and γ_k that play a significant role in determining this upper bound. Hence, in what follows, the analysis starts with a general form of α_k and γ_k, then a particular case is considered.

2.3.1 General Form of α_k and γ_k

We start by considering an additional assumption on the objective function for what follows.

Assumption 2.6. *Let $\mathcal{F}(x)$ be strongly convex, then there exists $\lambda > 0$ such that*

$$\langle \nabla \mathcal{F}(x), x - x^* \rangle \geq \lambda \|x - x^*\|^2.$$

Our main result regarding the convergence rate is summarized in the following theorem.

Theorem 2.7. *Let Assumptions 1.1-1.4, 2.1-2.2, and 2.6 hold. Define* $R = \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2$, $\delta_k = \left(\frac{1+\rho_w^2}{2}\right)^k$, *and* $\beta_k = \sum_{j=1}^{k} \delta_j \alpha_{k-j}^2$. *Let* $\bar{M}$ *denote the upper bound of* $\mathbb{E}[\|\bar{g}_k\|^2]$ *and* G *that of* $\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|$ *and define* $\bar{G} = \frac{2\rho_w^2}{1-\rho_w^2}G^2$ *(Refer to A, C.1, and C.2 for proof of boundedness). We have the following first result almost surely,*

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \delta_{k+1}R + \beta_{k+1}\bar{G} \tag{10}$$

(Refer to C.2 for proof).

Next, we define the constants $A = \lambda\alpha_2$, $B = \frac{2\alpha_2 L^2}{\lambda n}$, $C = \frac{\alpha_1^2\alpha_3^6}{2\alpha_2\lambda}$, *and*

$$K_0 = \arg\min_{A\alpha_k\gamma_k<1} k.$$

We finally define the following parameters:

$$\begin{aligned} \kappa_k &= \frac{1-(\frac{\gamma_{k+1}}{\gamma_k})^2}{\alpha_k\gamma_k}, \quad \sigma_1 = \max_{k\geq K_0} \kappa_k, \quad \sigma_2 = \max_{k\geq K_0} \frac{\delta_k}{\gamma_k^2}, \quad \sigma_3 = \max_{k\geq K_0} \frac{\beta_k}{\gamma_k^2}, \quad \sigma_4 = \max_{k\geq K_0} \frac{\alpha_k}{\gamma_k^3}, \\ \tau_k &= \frac{1-\frac{\alpha_{k+1}\gamma_{k+1}^{-1}}{\alpha_k\gamma_k^{-1}}}{\alpha_k\gamma_k}, \quad \sigma_5 = \max_{k\geq K_0} \tau_k, \quad \sigma_6 = \max_{k\geq K_0} \frac{\gamma_k^3}{\alpha_k}, \quad \sigma_7 = \max_{k\geq K_0} \frac{\gamma_k}{\alpha_k}\delta_k, \quad \sigma_8 = \max_{k\geq K_0} \frac{\gamma_k}{\alpha_k}\beta_k. \end{aligned} \tag{11}$$

If $\kappa_k < A$ *for any* $k \geq K_0$, *then*

$$D_k \leq \varsigma_1\gamma_k^2, \ \forall k \geq K_0, \tag{12}$$

with

$$\varsigma_1 \geq \max\left\{\frac{D_{K_0}}{\gamma_{K_0}^2}, \frac{BR\sigma_2 + B\bar{G}\sigma_3 + \bar{M}\sigma_4 + C}{A - \sigma_1}\right\}. \tag{13}$$

If $\tau_k < A$ *for any* $k \geq K_0$, *then*

$$D_k \leq \varsigma_2\frac{\alpha_k}{\gamma_k}, \ \forall k \geq K_0, \tag{14}$$

with

$$\varsigma_2 \geq \max\left\{\frac{D_{K_0}\gamma_{K_0}}{\alpha_{K_0}}, \frac{BR\sigma_7 + B\bar{G}\sigma_8 + C\sigma_6 + \bar{M}}{(A - \sigma_5)}\right\}. \tag{15}$$

Proof: See Appendix D.1.

2.3.2 A Special Case of α_k and γ_k

We now consider the special case mentioned in Example 2.3:

$$\alpha_k = \alpha_0(k+1)^{-\upsilon_1} \text{ and } \gamma_k = \gamma_0(k+1)^{-\upsilon_2}, \tag{16}$$

where $0.5 < \upsilon_1 < 1$ and $0 < \upsilon_2 \leq 1 - \upsilon_1$.

Before stating the main result, we consider the following lemma.

Lemma 2.8. *(Study of* β_k*) Let the step sizes have the form given in (16). Then, considering* K_1 *to be such that*

$$K_1 = \arg\min_{\left(\frac{1+\rho_w^2}{2}\right)^k \leq \alpha_k^2} k, \tag{17}$$

the convergence rate of β_k *and thus of* $\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2$ *is at least that of order* $\frac{1}{k^{3\upsilon_1-1}}$.

Proof: See Appendix D.4

We then let K_2 be such that $K_2 \geq \max\{K_0, K_1\}$ and state our next theorem.

Theorem 2.9. *Let α_k and γ_k have the forms given in (16) and consider the same assumptions of Theorem 2.7. If $\alpha_0\gamma_0 \geq \max\{2\upsilon_2, \upsilon_1 - \upsilon_2\}/A$, then we can say that there exists $\Upsilon < \infty$, where*

$$D_k \leq \Upsilon(k+1)^{-\min\{2\upsilon_2, \upsilon_1-\upsilon_2\}}, \ \forall k \geq K_2.$$

Proof: See Appendix D.5.

The parameters clearly affect the upper bound of the convergence rate or rate of expected divergence decay in Theorem 2.9. As it is evident that

$$\max\{2\upsilon_2, \upsilon_1 - \upsilon_2\} \leq 0.5,$$

the best choice is when equality holds for $\upsilon_1 = 0.75$ and $\upsilon_2 = 0.25$. With the sufficient condition on the parameters in Theorem 2.9, we can finally state that our algorithm converges with a rate of $O(\frac{1}{\sqrt{k}})$ after a sufficient number of iterations $k > K_2$ when the step sizes are vanishing.

2.4 Regret Bound

To further examine the performance of our algorithm, we present the following theorem on the achieved regret bound.

Theorem 2.10. *Let the assumptions of Theorem 2.7 hold. When α_k and γ_k have the forms of (16) with $\upsilon_1 = 0.75$ and $\upsilon_2 = 0.25$, the regret bound is given by*

$$\mathbb{E}_{\mathcal{H}_k}\left[\sum_{k=1}^{K} \mathcal{F}(\bar{x}_k) - \mathcal{F}(x^*)\right] \leq \Upsilon L(\sqrt{K+1} - 1).$$

Proof: See Appendix E.

2.5 Convergence Rate with Constant Step Sizes

In this subsection, we fix the step sizes to $\alpha_k = \alpha > 0$ and $\gamma_k = \gamma > 0$, $\forall k$, and we assume them to be two arbitrarily small values.

Theorem 2.11. *Assume $\alpha\gamma < \frac{1}{A}$ and $\alpha < \gamma$. Define $\varrho_1 = 1 - A\alpha\gamma$, $\varrho_2 = \frac{1+\rho_w^2}{2}$, and $\breve{G} = \frac{2\rho_w^2(1+\rho_w^2)G^2}{(1-\rho_w^2)^2}$. Let Assumptions 1.1-1.4, 2.2, and 2.6 hold, then*

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \varrho_2^{k+1} R + \alpha^2 \breve{G}. \tag{18}$$

Meaning, $\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2$ converges with the linear rate of $O\big(\varrho_2^k\big)$ for an arbitrary small α almost surely. Further,

- *When $\varrho_1 \leq \varrho_2$,*

$$D_{k+1} \leq \varrho_1^{k+1} D_0 + \varrho_2^{k+1} \frac{2\alpha\gamma BR}{2A\alpha\gamma + \rho_w^2 - 1} + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \tag{19}$$

  Then, for arbitrary small step sizes, D_k converges with the linear rate of $O\big(\varrho_2^k\big)$.

- *When $\varrho_1 > \varrho_2$,*

$$D_{k+1} \leq \varrho_1^{k+1}\Big(D_0 + \frac{2\alpha\gamma BR}{1 - 2A\alpha\gamma - \rho_w^2}\Big) + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \tag{20}$$

  Then, for arbitrary small step sizes, D_k converges with the linear rate of $O\big(\varrho_1^k\big)$.

Proof: See Appendix F.

Taking arbitrarily small values of α, γ satisfying $\alpha\gamma < \frac{1}{A}$ and $\alpha < \gamma$, the convergence rate becomes $O(\varrho^K)$, achieving the same rate as the gradient tracking technique with FO information.

3 Numerical Results

In this section, we provide numerical examples to illustrate the performance of the 1P-DSGT algorithm. We compare it with a general DSGT algorithm based on an unbiased estimator with bounded variance. For this unbiased estimator, we calculate the exact gradient and add white noise to it. The network topology is a connected Erdős-Rényi random graph with a probability of 0.05.

We consider a logistic classification problem to classify m images of the two digits, labeled as $y_{ij} = +1$ or -1 from the MNIST data set (LeCun & Cortes, 2005). Each image, X_{ij}, is a 785-dimensional vector and is compressed using a lossy autoencoder to become 10-dimensional, i.e., $d = 10$. The total images are split equally among the agents such that each agent has $m_i = \frac{m}{n}$ images and no access to other ones for privacy constraints. However, the goal is still to make use of all images and to solve collaboratively

$$min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m_i} \mathbb{E}_{u \sim \mathcal{N}(1,\sigma_u)} \ln(1 + \exp(-u_{ij} y_{ij} . X_{ij}^T \theta)) + c\|\theta\|^2,$$

while reaching consensus on the decision variable $\theta \in \mathbb{R}^d$. We note here that u models some perturbation on the local querying of every example to add to the randomization of the communication process.

We consider classifying the digits 1 and 2 with $m = 12700$ images. There are $n = 100$ agents in the network and thus each has a local batch of $m_i = 127$ images. We take $\sigma_u = 0.01$ and let $\alpha_k = 0.2(k+1)^{-0.75}$ and $\gamma_k = 0.7(k+1)^{-0.25}$ for 1P-DSGT with vanishing step sizes, and $\alpha = 0.1$ and $\gamma = 0.7$ with constant step sizes. We choose $\Phi_k \in \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$ with equal probability. Also, every function query is subject to a white noise generated by the standard normal distribution. For the general DSGT algorithm, we set the step size to $\alpha_k = 0.015(k+1)^{-1}$ when it is vanishing and $\alpha = 0.015$ when constant, and we do not consider the perturbation on the objective function nor the noise on the objective function, only the noise on the exact gradient. We let $c = 0.1$, and the initialization be the same for both algorithms, with $\theta_{i,0}$ uniformly chosen from $[-0.5, 0.5]^d$, $\forall i \in \mathcal{N}$, per instance. We finally average the simulations over 30 instances.

The expected evolution of the loss objective function is presented in Figure 1 and the graphs are zoomed in on in Figure 2. Experimental results seem to validate our theoretical results: 1P-DSGT converges linearly fast with constant step sizes, however the final gap is due to converging to an $O(\alpha)$-neighborhood of the optimal solution. 1P-DSGT with vanishing step sizes converges with an $O(\frac{1}{\sqrt{k}})$ while DSGT with vanishing step size converges at a rate of $O(\frac{1}{k})$ (Pu & Nedić, 2018). Using constant vs vanishing step size does not seem to affect the convergence rate of the loss function of DSGT.

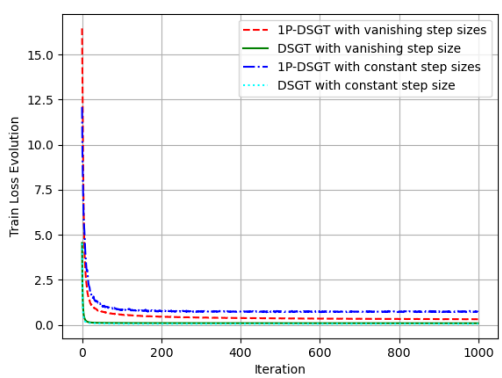

Figure 1: Expected loss function evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

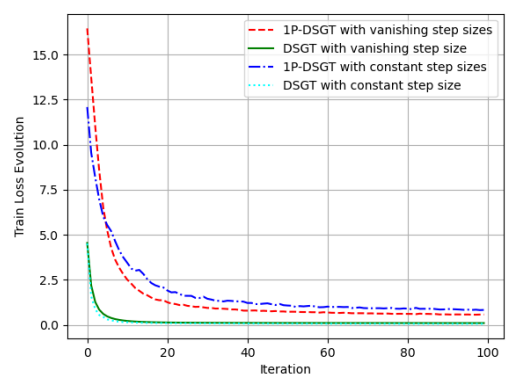

Figure 2: Expected loss function evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

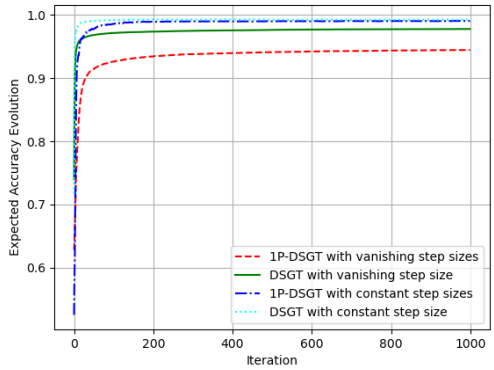

Figure 3: Expected test accuracy evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

In Figure 3, we measure at every iteration the classification accuracy against an independent test set of 2167 images using the updated mean vector $\bar{\theta}_k = \frac{1}{n}\sum_{i=1}^{n}\theta_{i,k}$ of the local decision variables. The interest of the constant step sizes appears in the convergence rate of this accuracy, where our algorithm 1P-DSGT is able to compete with DSGT with full FO information, and to outperform DSGT with a vanishing step size. This is an important result as it shows that the classification goal with ZO is well met despite the limiting upper bounds of convergence rate and that $O(\alpha)$-neighborhood of the optimal solution achieved linearly fast can be sufficient to achieve the best possible accuracy.

In Figures 4 and 5, the curves are those of the evolution of the expected consensus error, or $\mathbb{E}\big[\sum_{i=1}^{n}\|\theta_{i,k} - \bar{\theta}_k\|^2\big]$ which is the expected error between the local decision variables and their average. For both algorithms, the error again validates the theoretical bounds and decreases quite fast.

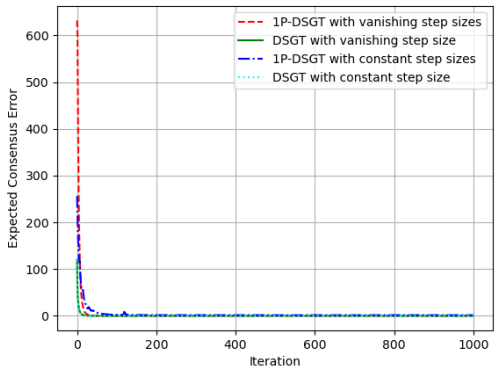

Figure 4: Expected consensus error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

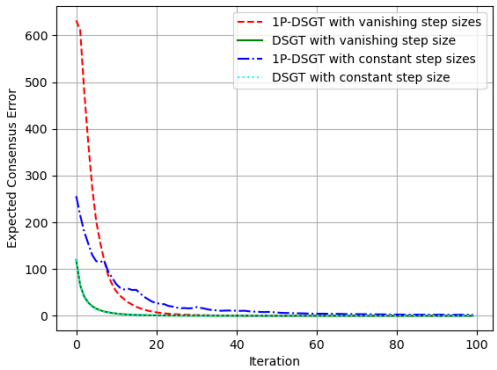

Figure 5: Expected consensus error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

Figures 6 and 7 show the evolution of the expected gradient tracking error, or $\mathbb{E}\big[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\big]$ which is the expected error between the auxiliary local variables and their average. Despite not knowing the theoretical constants that bound the tracking error of both algorithms, we see that the final error attained by the ZO gradient is much smaller than that of the FO one no matter the step sizes. This shows in a sense that the ZO gradient is "easier to track" across network agents.

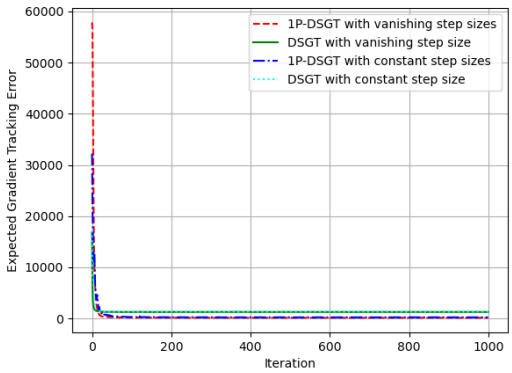

Figure 6: Expected gradient tracking error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

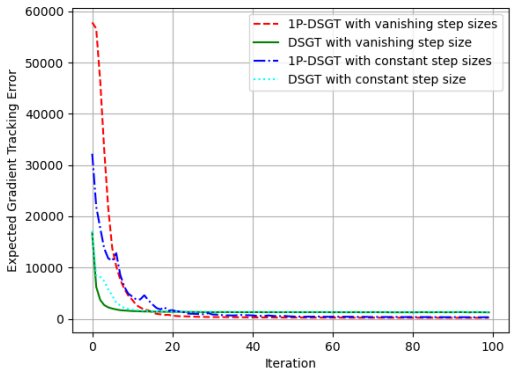

Figure 7: Expected gradient tracking error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes.

We add other numerical examples for different image labels in Appendix G.

4 Conclusion

In this work, we extended the gradient-tracking algorithm to present a practical solution to a relevant problem with realistic assumptions. A distributed stochastic gradient-tracking algorithm was studied and proved to converge with a biased and high variance one-point gradient estimate and a stochastic perturbation on the objective function. In the context of noisy ZO optimization, we have successfully established a linear convergence rate of $O(\varrho^k)$ using fixed step sizes and $O(\frac{1}{\sqrt{k}})$ with vanishing step sizes. These rates align with the optimal expectations examined in the existing literature. We also prove a regret bound that of $O(\sqrt{k})$ with vanishing step sizes. A numerical application confirmed the success and efficiency of the algorithm.

References

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, 2010.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization vianney perchet, 2016. URL `https://arxiv.org/abs/1605.08165`.

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *J. ACM*, 68(4), jun 2021. ISSN 0004-5411. doi: 10.1145/3453721. URL `https://doi.org/10.1145/3453721`.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *NeurIPS*, 2019.

Ying Chen, Zhiyong Liu, Yongchao Zhang, Yuan Wu, Xin Chen, and Lian Zhao. Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things. *IEEE Transactions on Industrial Informatics*, 17(7):4925–4934, 2021. doi: 10.1109/TII.2020.3028963.

Amir Daneshmand, Francisco Facchinei, Vyacheslav Kungurtsev, and Gesualdo Scutari. Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization. *IEEE Transactions on Signal Processing*, 63(15):3914–3929, 2015. doi: 10.1109/TSP.2015.2436357.

Joseph L. Doob. Stochastic processes. 1953.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015. doi: 10.1109/TIT.2015.2409256.

Abraham Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *CoRR*, cs.LG/0408007, 2004. URL http://arxiv.org/abs/cs.LG/0408007.

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi (Richard) Zhang. Gradientless descent: High-dimensional zeroth-order optimization. *CoRR*, abs/1911.06317, 2019. URL http://arxiv.org/abs/1911.06317.

Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization, 2018. URL https://arxiv.org/abs/1802.09022.

Davood Hajinezhad, Mingyi Hong, and Alfredo Garcia. Zone: Zeroth-order nonconvex multiagent optimization over networks. *IEEE Transactions on Automatic Control*, 64(10):3995–4010, 2019. doi: 10.1109/TAC.2019.2896025.

Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *NIPS*, 2012.

Xia Jiang, Xianlin Zeng, Jian Sun, and Jie Chen. Distributed stochastic gradient tracking algorithm with variance reduction for non-convex optimization. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022. doi: 10.1109/TNNLS.2022.3170944.

Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic, and Soummya Kar. Distributed zeroth order optimization over random networks: A kiefer-wolfowitz stochastic approximation approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 4951–4958, 2018. doi: 10.1109/CDC.2018.8619044.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

Huaqing Li, Lifeng Zheng, Zheng Wang, Yu Yan, Liping Feng, and Jing Guo. S-diging: A stochastic gradient tracking algorithm for distributed optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):53–65, 2022. doi: 10.1109/TETCI.2020.3017242.

Wenjie Li and Mohamad Assaad. Distributed stochastic optimization in networks with low informational exchange. *IEEE Transactions on Information Theory*, 67(5):2989–3008, 2021. doi: 10.1109/TIT.2021.3064925.

Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 288–297. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/liu18a.html.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *ICLR*, 2019.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O. Hero III, and Pramod K. Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020. doi: 10.1109/MSP.2020.3003837.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. doi: 10.1109/TSIPN.2016.2524588.

Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: a gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pp. 315–321, 2019. doi: 10.1109/DSW.2019.8755807.

Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief. A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys Tutorials*, 19(4):2322–2358, 2017. doi: 10.1109/COMST.2017.2745201.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017. doi: 10.1137/16M1084316. URL https://doi.org/10.1137/16M1084316.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Yurii Nesterov and Vladimir G. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Shi Pu. A robust gradient tracking method for distributed optimization over directed networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2335–2341, 2020. doi: 10.1109/CDC42340.2020.9303917.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods, 2018. URL https://arxiv.org/abs/1805.11454.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018. doi: 10.1109/TCNS.2017.2698261.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 3–24, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Shamir13.html.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015. doi: 10.1137/14096668X. URL https://doi.org/10.1137/14096668X.

Zai Shi and Atilla Eryilmaz. A zeroth-order admm algorithm for stochastic optimization over distributed processing networks. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, pp. 726–735, 2020. doi: 10.1109/INFOCOM41043.2020.9155520.

Yujie Tang, Junshan Zhang, and Na Li. Distributed zero-order algorithms for nonconvex multiagent optimization. *IEEE Transactions on Control of Network Systems*, 8(1):269–281, 2021. doi: 10.1109/TCNS.2020.3024321.

Anirudh Vemula, Wen Sun, and J. Andrew Bagnell. Contrasting exploration in parameter and action space: A zeroth-order optimization perspective. *ArXiv*, abs/1901.11503, 2019.

Ran Xin, Anit Kumar Sahu, Usman A. Khan, and Soummya Kar. Distributed stochastic optimization with gradient tracking over strongly-connected networks. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 8353–8358, 2019. doi: 10.1109/CDC40024.2019.9029217.

Fanqin Zhou, Lei Feng, Michel Kadoch, Peng Yu, Wenjing Li, and Zhili Wang. Multiagent rl aided task offloading and resource management in wi-fi 6 and 5g coexisting industrial wireless environment. *IEEE Transactions on Industrial Informatics*, 18(5):2923–2933, 2022. doi: 10.1109/TII.2021.3106973.

A Estimated Gradient

In this section, we derive the bias of the gradient estimate with respect to the real gradient of the local objective function. Let

$$\breve{g}_{i,k} = \mathbb{E}_{S,\Phi,\zeta}[g_{i,k}|\mathcal{H}_k].$$

Thus, by Assumption 1.4 and the definition in (4),

$$\begin{aligned}\breve{g}_{i,k} &= \mathbb{E}_{S,\Phi,\zeta}[\Phi_{i,k}(f_i(x_{i,k}+\gamma_k\Phi_{i,k}, S_{i,k})+\zeta_{i,k})|\mathcal{H}_k] \\ &= \mathbb{E}_{S,\Phi}[\Phi_{i,k} f_i(x_{i,k}+\gamma_k\Phi_{i,k}, S_{i,k})|\mathcal{H}_k] \\ &= \mathbb{E}_{\Phi}[\Phi_{i,k} F_i(x_{i,k}+\gamma_k\Phi_{i,k})|\mathcal{H}_k].\end{aligned}$$

By Taylor's theorem and the mean-valued theorem, there exists $\tilde{x}_{i,k}$ located between $x_{i,k}$ and $x_{i,k}+\gamma_k\Phi_{i,k}$ where

$$F_i(x_{i,k}+\gamma_k\Phi_{i,k}) = F_i(x_{i,k}) + \gamma_k\langle\Phi_{i,k}, \nabla F_i(x_{i,k})\rangle + \frac{\gamma_k^2}{2}\langle\Phi_{i,k}, \nabla^2 F_i(\tilde{x}_{i,k})\Phi_{i,k}\rangle,$$

substituting in the previous definition,

$$\begin{aligned}\breve{g}_{i,k} &= F_i(x_{i,k})\mathbb{E}_\Phi[\Phi_{i,k}] + \gamma_k\mathbb{E}_\Phi[\Phi_{i,k}\Phi_{i,k}^T]\nabla F_i(x_{i,k}) + \frac{\gamma_k^2}{2}\mathbb{E}_\Phi[\Phi_{i,k}\Phi_{i,k}^T\nabla^2 F_i(\tilde{x}_{i,k})\Phi_{i,k}|\mathcal{H}_k] \\ &= \alpha_2\gamma_k[\nabla F_i(x_{i,k}) + b_{i,k}].\end{aligned}$$

Thus, the estimation bias has the form

$$\begin{aligned}b_{i,k} &= \frac{\breve{g}_{i,k}}{\alpha_2\gamma_k} - \nabla F_i(x_{i,k}) \\ &= \frac{\gamma_k}{2\alpha_2}\mathbb{E}_\Phi[\Phi_{i,k}\Phi_{i,k}^T\nabla^2 F_i(\tilde{x}_{i,k})\Phi_{i,k}|\mathcal{H}_k].\end{aligned}$$

Let Assumptions 1.2 and 2.2 hold. Then, we can bound the bias as

$$\begin{aligned}\|b_{i,k}\| &\le \frac{\gamma_k}{2\alpha_2}\mathbb{E}_\Phi[\|\Phi_{i,k}\|_2\|\Phi_{i,k}^T\|_2\|\nabla^2 F_i(\tilde{x}_{i,k})\|_2\|\Phi_{i,k}\|_2|\mathcal{H}_k] \\ &\le \gamma_k\frac{\alpha_3^3\alpha_1}{2\alpha_2}.\end{aligned}$$

We can see $\|b_{i,k}\| \to 0$ as $k \to \infty$ since γ_k is vanishing. We remark that

$$\begin{aligned}\tilde{g}_k &= \mathbb{E}[\bar{g}_k|\mathcal{H}_k] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[g_{i,k}|\mathcal{H}_k] \\ &= \frac{1}{n}\sum_{i=1}^n \alpha_2\gamma_k[\nabla F_i(x_{i,k}) + b_{i,k}] \\ &= \alpha_2\gamma_k[h(\mathbf{x}_k) + \bar{b}_k]\end{aligned} \tag{21}$$

is also a biased estimator of $h(\mathbf{x}_k)$ with

$$\begin{aligned}\|\bar{b}_k\| &= \|\frac{1}{n}\sum_{i=1}^n b_{i,k}\| \\ &\le \frac{1}{n}\sum_{i=1}^n \|b_{i,k}\| \\ &\le \frac{1}{n}\sum_{i=1}^n \gamma_k\frac{\alpha_3^3\alpha_1}{2\alpha_2} \\ &= \gamma_k\frac{\alpha_3^3\alpha_1}{2\alpha_2}.\end{aligned} \tag{22}$$

Lemma A.1. *Let all Assumptions 1.3, 1.4, and 2.2 hold and $\|\mathbf{x}_k\| < \infty$ almost surely, then there exists a bounded constant $M > 0$, such that $E[\|\mathbf{g}_k\|^2] < M$ almost surely.*

Proof. $\forall i \in \mathcal{N}$, we have

$$\begin{aligned}
\mathbb{E}[\|g_{i,k}\|^2|\mathcal{H}_k] &= \mathbb{E}[\|\Phi_{i,k}(f_i(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k}) + \zeta_{i,k})\|^2|\mathcal{H}_k] \\
&= \mathbb{E}[\|\Phi_{i,k}\|^2\|f_i(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k}) + \zeta_{i,k}\|^2|\mathcal{H}_k] \\
&\overset{(a)}{\leq} \alpha_3^2\mathbb{E}[(f_i(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k}) + \zeta_{i,k})^2|\mathcal{H}_k] \\
&\overset{(b)}{=} \alpha_3^2\mathbb{E}[f_i^2(x_{i,k} + \gamma_k\Phi_{i,k}, S_{i,k})|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\
&\overset{(c)}{\leq} \alpha_3^2\mathbb{E}[(\|f_i(0, S_{i,k})\| + L_{S_{i,k}}\|x_{i,k} + \gamma_k\Phi_{i,k}\|)^2|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\
&\overset{(d)}{\leq} 2\alpha_3^2\mathbb{E}[\mu_{S_{i,k}}^2 + L_{S_{i,k}}^2(\|x_{i,k}\| + \gamma_k\alpha_3)^2|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\
&\overset{(e)}{=} 2\alpha_3^2(\mu + L'(\|x_{i,k}\| + \gamma_k\alpha_3)^2) + \alpha_3^2\alpha_4 \\
&< \infty,
\end{aligned}$$

where (a) is due to Assumption 2.2, (b) Assumption 1.4, and (c) Assumption 1.3. We denote $\|f_i(0, S_{i,k})\| = \mu_{S_{i,k}}$ in (d) and the inequality is due to $\frac{x+y}{2} \leq \sqrt{\frac{x^2+y^2}{2}}$, $\forall x, y \in \mathbb{R}$. In (e), $\mu = \mathbb{E}[\mu_{S_{i,k}}^2]$ and $L' = \mathbb{E}[L_{S_{i,k}}^2]$. □

B Stochastic Noise

To prove Lemma 2.4, we begin by demonstrating that the sequence $\{\sum_{k=K}^{K'} \alpha_k e_k\}_{K' \geq K}$ is a martingale. To do so, we have to prove that for all $K' \geq K$, $X_{K'} = \sum_{k=K}^{K'} \alpha_k e_k$ satisfies the following two conditions:

(i) $\mathbb{E}[X_{K'+1}|X_{K'}] = X_{K'}$

(ii) $\mathbb{E}[\|X_{K'}\|^2] < \infty$

We know that

$$\mathbb{E}[e_k] = \mathbb{E}[\bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k]] = \mathbb{E}_{\mathcal{H}_k}\Big[\mathbb{E}\Big[\bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\Big|\mathcal{H}_k\Big]\Big] = 0$$

by the law of total expectation. Hence,

$$\mathbb{E}[X_{K'+1}|X_{K'}] = \mathbb{E}\Big[\alpha_{K'+1}e_{K'+1} + \sum_{k=K}^{K'} \alpha_k e_k\Big|\sum_{k=K}^{K'} \alpha_k e_k\Big] = 0 + \sum_{k=K}^{K'} \alpha_k e_k = X_{K'}. \tag{23}$$

In addition, e_k and $e_{k'}$ are uncorrelated for any $k \neq k'$ since (assuming $k > k'$) $\mathbb{E}\big[e_k^T e_{k'}\big] = \mathbb{E}\big[\mathbb{E}[e_k^T e_{k'}|\mathcal{H}_k]\big] = \mathbb{E}\big[e_{k'}\mathbb{E}[e_k^T|\mathcal{H}_k]\big] = 0$. Thus,

$$\begin{aligned}
\mathbb{E}(\|\sum_{k=K}^{K'} \alpha_k e_k\|^2) &= \mathbb{E}(\sum_{k=K}^{K'}\sum_{k'=K}^{K'} \alpha_k \alpha_{k'} \langle e_k, e_{k'}\rangle) \\
&\overset{(a)}{=} \mathbb{E}(\sum_{k=K}^{K'} \|\alpha_k e_k\|^2) \\
&\leq \sum_{k=K}^{\infty} \mathbb{E}(\alpha_k^2 \|\bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\|^2) \\
&= \sum_{k=K}^{\infty} \alpha_k^2 \mathbb{E}(\|\bar{g}_k\|^2) - \mathbb{E}_{\mathcal{H}_k}(\|\mathbb{E}[\bar{g}_k|\mathcal{H}_k]\|^2) \\
&\leq \sum_{k=K}^{\infty} \alpha_k^2 \mathbb{E}(\|\bar{g}_k\|^2) \\
&\overset{(b)}{\leq} M \sum_{k=K}^{\infty} \alpha_k^2 \\
&\overset{(c)}{<} \infty,
\end{aligned} \tag{24}$$

where (a) is due to the uncorrelatedness $\mathbb{E}[\langle e_k, e_{k'}\rangle] = 0$, (b) is by Lemma A.1, and (c) is by Assumption 2.1. Therefore, both (i) and (ii) are satisfied and we can say that $\{\sum_{k=K}^{K'} \alpha_k e_k\}_{K' \geq K}$ is a martingale. This permits us to use Doob's martingale inequality Doob (1953):

For any constant $\nu > 0$,

$$\begin{aligned}
\mathbb{P}(\sup_{K' \geq K} \|\sum_{k=K}^{K'} \alpha_k e_k\| \geq \nu) &\leq \frac{1}{\nu^2}\mathbb{E}(\|\sum_{k=K}^{K'} \alpha_k e_k\|^2) \\
&\overset{(a)}{\leq} \frac{M}{\nu^2}\sum_{k=K}^{\infty} \alpha_k^2,
\end{aligned} \tag{25}$$

where (a) is following the exact same steps as (24).

Since M is a bounded constant and $\lim_{K\to\infty} \sum_{k=K}^{\infty} \alpha_k^2 = 0$ by Assumption 2.1, we get $\lim_{K\to\infty} \frac{M}{\nu^2} \sum_{k=K}^{\infty} \alpha_k^2 = 0$ for any bounded constant ν. Hence, the probability that $\|\sum_{k=K}^{K'} \alpha_k e_k\| \geq \nu$ also vanishes as $K \to \infty$, which concludes the proof.

C Proof of Convergence

We start by stating the following lemma that will be useful for the proof of convergence.

Lemma C.1. *If all Assumptions 1.1-1.4 and 2.1-2.2 hold and* $\|\mathbf{x}_k\| < \infty$ *almost surely, then* $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$. *In fact, we have* $\sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 < \infty$ *as well as*

$$\sum_{k=0}^{\infty} \gamma_k \alpha_k \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| < \infty$$

almost surely.

Proof: See C.2.

C.1 Proof of Theorem 2.5

The goal is to write the divergence in terms of its previous term and to prove that it is finally vanishing. We know that $\bar{x}_{k+1} = \bar{x}_k - \alpha_k \bar{g}_k$. With this equation, the divergence at time $k+1$ can be written as

$$\begin{aligned}
&d_{k+1}\\
=&\|\bar{x}_{k+1} - x^*\|^2\\
=&\|\bar{x}_k - \alpha_k \bar{g}_k - x^*\|\\
=&\|\bar{x}_k - x^*\|^2 - 2\alpha_k\langle \bar{x}_k - x^*, \bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k] + \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle + \alpha_k^2\|\bar{g}_k\|^2\\
=&\|\bar{x}_k - x^*\|^2 - 2\alpha_k\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle - 2\alpha_k\langle \bar{x}_k - x^*, e_k\rangle + \alpha_k^2\|\bar{g}_k\|^2\\
\overset{(a)}{=}&d_k - 2\alpha_2\gamma_k\alpha_k\langle \bar{x}_k - x^*, h(\mathbf{x}_k) + \bar{b}_k\rangle - 2\alpha_k\langle \bar{x}_k - x^*, e_k\rangle + \alpha_k^2\|\bar{g}_k\|^2\\
=&d_k - 2\alpha_2\gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle + 2\alpha_2\gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle - 2\alpha_k\langle \bar{x}_k - x^*, e_k\rangle + \alpha_k^2\|\bar{g}_k\|^2\\
\overset{(b)}{\leq}&d_k - 2\alpha_2\gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle + \frac{2\alpha_2 L\gamma_k\alpha_k}{\sqrt{n}}\|\bar{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| - 2\alpha_k\langle \bar{x}_k - x^*, e_k\rangle + \alpha_k^2\|\bar{g}_k\|^2,
\end{aligned}$$

where (a) is due to (21) and (b) is due to Lemma 1.6. By recursion, we have

$$\begin{aligned}
d_{K+1} \leq d_0 - 2\alpha_2\sum_{k=0}^{K}\gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle + \frac{2\alpha_2 L}{\sqrt{n}}\sum_{k=0}^{K}\gamma_k\alpha_k\|\bar{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|\\
- 2\sum_{k=0}^{K}\alpha_k\langle \bar{x}_k - x^*, e_k\rangle + \sum_{k=0}^{K}\alpha_k^2\|\bar{g}_k\|^2.
\end{aligned} \tag{26}$$

By Lemma 2.4, we have $\lim_{K\to\infty}\|\sum_{k=0}^{K}\alpha_k e_k\| < \infty$ almost surely. Since $\|\bar{x}_k - x^*\| = \|\frac{1}{n}\sum_{i=1}^{n}(x_{i,k} - x^*)\| \leq \frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - x^*\| < \infty$ almost surely, hence

$$\lim_{K\to\infty}\|\sum_{k=0}^{K}\alpha_k\langle \bar{x}_k - x^*, e_k\rangle\| < \infty. \tag{27}$$

From Lemma A.1, we have

$$\|\bar{g}_k\|^2 = \|\frac{1}{n}\sum_{i=1}^{n}g_{i,k}\|^2 \leq \frac{n}{n^2}\sum_{i=1}^{n}\|g_{i,k}\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|g_{i,k}\|^2 < \infty, \text{ almost surely.} \tag{28}$$

Then, by Assumption 2.1,

$$\lim_{K\to\infty}\sum_{k=0}^{K}\alpha_k^2\|\bar{g}_k\|^2 < \infty. \tag{29}$$

As stated in Lemma C.1, we have $\sum_{k=0}^{\infty}\gamma_k\alpha_k\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| < \infty$, adding to $\|\bar{x}_k - x^*\| < \infty$ almost surely, then

$$\lim_{K\to\infty}\sum_{k=0}^{K}\gamma_k\alpha_k\|\bar{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| < \infty. \tag{30}$$

From the above inequalities (26)-(30), we conclude that there exists D such that $d_{K+1} \leq D + z_K$, with

$$z_K = -2\alpha_2\sum_{k=0}^{K}\gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle. \tag{31}$$

Carefully examine (22), we can say that there exists K_b, such that for $k \geq K_b$, $\bar{b}_k$ becomes very small where we can find an arbitrary small positive value ϵ_b such that,

$$\|\nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\| \geq (1 - \epsilon_b)\|\nabla\mathcal{F}(\bar{x}_k)\|, \ \forall k \geq K_b, \tag{32}$$

leading to $-\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle \leq 0$, due to the convexity of $\mathcal{F}$ in (5).

Consequently, for any big K, $0 \leq d_{K+1} < \infty$ and the limit $\lim_{K\to\infty} d_{K+1} = \bar{d}$ exists.

Thus, there are 2 cases: $\bar{d} > 0$ or $\bar{d} = 0$. Assume hypothesis *H1*) $\bar{d} > 0$ to be valid, i.e., $\bar{x}_k$ does not converge to x^*, then $\forall \epsilon_h > 0$, $\exists K_h$ such that

$$-\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k)\rangle < -\epsilon_h,\ \forall k \geq K_h.$$

From (31) and (32), we get that $\forall k \geq K_m = \max\{K_b, K_h\}$,

$$-\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle < -\epsilon_h(1-\epsilon_b),$$

implying

$$\lim_{K\to\infty} -\sum_{k=K_m}^{K} \gamma_k\alpha_k\langle \bar{x}_k - x^*, \nabla\mathcal{F}(\bar{x}_k) + \bar{b}_k\rangle < -\epsilon_h(1-\epsilon_b)\lim_{K\to\infty}\sum_{k=K_m}^{K}\gamma_k\alpha_k < -\infty$$

since $\sum \gamma_k\alpha_k$ diverges by Assumption 2.1. As a result, we get $z_K < -\infty$ and $d_{K+1} < -\infty$. However, by definition in (9), $d_{K+1} > 0$. Accordingly, the hypothesis *H1* cannot be true and the case $\bar{d} = 0$ is the valid one. We conclude that $\lim_{k\to\infty} d_k = 0$, $\lim_{k\to\infty}\nabla\mathcal{F}(\bar{x}_k) = 0$, and $\lim_{k\to\infty}\bar{x}_k = x^*$ almost surely.

C.2 Proof of Lemma C.1

We start by replacing the variables with their algorithmic updates.

$$\begin{aligned}
&\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2\\
=&\|W\mathbf{x}_k - \alpha_k W\mathbf{y}_k - \mathbf{1}\bar{x}_k + \alpha_k\mathbf{1}\bar{y}_k\|^2\\
=&\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 - 2\alpha_k\langle W\mathbf{x}_k - \mathbf{1}\bar{x}_k, W\mathbf{y}_k - \mathbf{1}\bar{y}_k\rangle + \alpha_k^2\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\\
\overset{(a)}{\leq}&\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \alpha_k[\frac{1-\rho_w^2}{2\rho_w^2\alpha_k}\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2\alpha_k}{1-\rho_w^2}\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \alpha_k^2\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\\
\overset{(b)}{\leq}&\rho_w^2\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \rho_w^2\alpha_k[\frac{1-\rho_w^2}{2\rho_w^2\alpha_k}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2\alpha_k}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \rho_w^2\alpha_k^2\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\\
=&\frac{1+\rho_w^2}{2}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \alpha_k^2\frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2
\end{aligned} \tag{33}$$

where (a) is by $-2\epsilon \times \frac{1}{\epsilon}\langle a, b\rangle = -2\langle \epsilon a, \frac{1}{\epsilon}b\rangle \leq \epsilon^2\|a\|^2 + \frac{1}{\epsilon^2}\|b\|^2$ and (b) is by Lemma 1.5. By induction, we have

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \big(\frac{1+\rho_w^2}{2}\big)^{k+1}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \frac{2\rho_w^2}{1-\rho_w^2}\sum_{j=0}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{j+1}\alpha_{k-j}^2\|\mathbf{y}_{k-j} - \mathbf{1}\bar{y}_{k-j}\|^2. \tag{34}$$

Since $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$,

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\| \leq \big(\frac{1+\rho_w^2}{2}\big)^{\frac{k+1}{2}}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\| + \sqrt{\frac{2\rho_w^2}{1-\rho_w^2}}\sum_{j=0}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-j}\|\mathbf{y}_{k-j} - \mathbf{1}\bar{y}_{k-j}\|. \tag{35}$$

By repeatedly replacing the variables with the algorithm's iterations, we see that

$$
\begin{aligned}
\mathbf{y}_{k+1} &= W\mathbf{y}_k + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W(W\mathbf{y}_{k-1} + \mathbf{g}_k - \mathbf{g}_{k-1}) + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W^2\mathbf{y}_{k-1} - W\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^2(W\mathbf{y}_{k-2} + \mathbf{g}_{k-1} - \mathbf{g}_{k-2}) - W\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^3\mathbf{y}_{k-2} - W^2\mathbf{g}_{k-2} + W(W-I)\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^3(W\mathbf{y}_{k-3} + \mathbf{g}_{k-2} - \mathbf{g}_{k-3}) - W^2\mathbf{g}_{k-2} + W(W-I)\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^4\mathbf{y}_{k-3} - W^3\mathbf{g}_{k-3} + W^2(W-I)\mathbf{g}_{k-2} + W(W-I)\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= \dots \\
&= W^{k+1}\mathbf{y}_0 - W^k\mathbf{g}_0 + \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= W^k(W-I)\mathbf{g}_0 + \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= \sum_{j=0}^{k} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1},
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{y}_k - \mathbf{1}\bar{y}_k &= \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1} \frac{1}{n}\mathbf{1}\mathbf{1}^T W^j(W-I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1} \frac{1}{n}\mathbf{1}\mathbf{1}^T(W-I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1} (W^j - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1} (W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \mathbf{1}\bar{g}_k,
\end{aligned}
$$

where the last equality can be proven by recursion and the fact that the matrix W is doubly stochastic by Assumption 1.1:

$(W-\frac{1}{n}\mathbf{1}\mathbf{1}^T)^{j+1} = (W^j-\frac{1}{n}\mathbf{1}\mathbf{1}^T)(W-\frac{1}{n}\mathbf{1}\mathbf{1}^T) = W^{j+1}-\frac{1}{n}W^j\mathbf{1}\mathbf{1}^T-\frac{1}{n}\mathbf{1}\mathbf{1}^TW+\frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1}-\frac{2}{n}\mathbf{1}\mathbf{1}^T+\frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1}-\frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Thus,

$$
\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \sum_{j=0}^{k-1} \|(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j(W-I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\| \\
&\leq \sum_{j=0}^{k-1} \rho_w^j\|(W-I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|.
\end{aligned}
$$

From Lemma A.1, we have $\|\mathbf{g}_k\|^2 < \infty$ almost surely.

$$\begin{aligned}
\|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|^2 &= \sum_{i=1}^{n} \|g_{i,k} - \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\|^2 \\
&= \sum_{i=1}^{n} \left(\|g_{i,k}\|^2 - 2\langle g_{i,k}, \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\rangle + \|\bar{g}_k\|^2 \right) \\
&= \|\mathbf{g}_k\|^2 - 2n\|\bar{g}_k\|^2 + n\|\bar{g}_k\|^2 \\
&= \|\mathbf{g}_k\|^2 - n\|\bar{g}_k\|^2 \\
&\leq \|\mathbf{g}_k\|^2 \\
&\leq M'^2 < \infty.
\end{aligned}$$

Inserting in the previous inequality, we get

$$\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \frac{M'}{1-\rho_w}\|(W-I)\| + M' \\
&= G < \infty,
\end{aligned} \tag{36}$$

where we have a geometric sum as $\rho_w < 1$.

1. **Proving** $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$
   Reconsider (33),

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &\leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \alpha_k^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\
\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 &\leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_{k-1} - \mathbf{1}\bar{x}_{k-1}\|^2 + \alpha_{k-1}^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_{k-1} - \mathbf{1}\bar{y}_{k-1}\|^2 \\
&\dots \\
\|\mathbf{x}_1 - \mathbf{1}\bar{x}_1\|^2 &\leq \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \alpha_0^2 \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_0 - \mathbf{1}\bar{y}_0\|^2.
\end{aligned} \tag{37}$$

   Adding all inequalities in (37), we obtain

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &\leq -\frac{1-\rho_w^2}{2}\sum_{i=1}^{k}\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2 + \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 \\
&\quad + \frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\sum_{i=0}^{k}\alpha_i^2\|\mathbf{y}_i - \mathbf{1}\bar{y}_i\|^2 \\
&\overset{(a)}{\leq} -\frac{1-\rho_w^2}{2}\sum_{i=1}^{k}\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2 + \frac{1+\rho_w^2}{2}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 \\
&\quad + G^2\frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\sum_{i=0}^{k}\alpha_i^2,
\end{aligned}$$

   with (a) being due to (36). Let $k \to \infty$, then the second and third terms are bounded due to Assumption 2.1. There are then 2 cases: $\sum\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ either diverges or converges. Assume the validity of the hypothesis *H2*) $\sum\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ diverges, i.e., $\sum_{i=1}^{\infty}\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2 \to \infty$. This leads to

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 < -\infty,$$

   as $-\frac{1-\rho_w^2}{2} < 0$. However, $\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2$ should be positive. Thus, hypothesis *H2* cannot be true and $\sum\|\mathbf{x}_i - \mathbf{1}\bar{x}_i\|^2$ converges. Hence, $\lim_{k\to\infty}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 = 0$ almost surely.

2. **Proving** $\sum_{k=0}^{\infty}\gamma_k\alpha_k\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|<\infty$

   Going back to (35),

$$\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\|\leq\big(\frac{1+\rho_w^2}{2}\big)^{\frac{k+1}{2}}\|\mathbf{x}_0-\mathbf{1}\bar{x}_0\|+G\sqrt{\frac{2\rho_w^2}{1-\rho_w^2}}\sum_{j=0}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-j},$$

   then substituting into the sum $\sum_{k=0}^{\infty}\gamma_k\alpha_k\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|$,

$$\begin{aligned}&\sum_{k=1}^{\infty}\gamma_k\alpha_k\left(\big(\frac{1+\rho_w^2}{2}\big)^{\frac{k}{2}}\|\mathbf{x}_0-\mathbf{1}\bar{x}_0\|+G\sqrt{\frac{2\rho_w^2}{1-\rho_w^2}}\sum_{j=0}^{k-1}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-1-j}\right)\\ \leq&\gamma_0\alpha_0\|\mathbf{x}_0-\mathbf{1}\bar{x}_0\|\frac{\sqrt{1+\rho_w^2}}{\sqrt{2}-\sqrt{1+\rho_w^2}}+G\sqrt{\frac{2\rho_w^2}{1-\rho_w^2}}\sum_{k=1}^{\infty}\gamma_k\alpha_k\sum_{j=0}^{k-1}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-1-j},\end{aligned}$$

   where the inequality is due to the fact that γ_k and α_k are both decreasing step-sizes and we have a geometric sum of ratio $\sqrt{\frac{1+\rho^2}{2}}<1$. We then study the sums in the second term,

$$\begin{aligned}\sum_{k=1}^{\infty}\gamma_k\alpha_k\sum_{j=0}^{k-1}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-1-j}\leq&\sum_{k=1}^{\infty}\gamma_k\sum_{j=0}^{k-1}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{j+1}{2}}\alpha_{k-1-j}^2\\ =&\sum_{k=1}^{\infty}\gamma_k\sum_{j=1}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{k-j+1}{2}}\alpha_{j-1}^2\\ =&\sum_{j=1}^{\infty}\alpha_{j-1}^2\sum_{k=j}^{\infty}\gamma_k\big(\frac{1+\rho_w^2}{2}\big)^{\frac{k-j+1}{2}}\\ \leq&\gamma_0\sum_{j=1}^{\infty}\alpha_{j-1}^2\sum_{k=j}^{\infty}\big(\frac{1+\rho_w^2}{2}\big)^{\frac{k-j+1}{2}}\\ =&\gamma_0\frac{\sqrt{1+\rho_w^2}}{\sqrt{2}-\sqrt{1+\rho_w^2}}\sum_{j=1}^{\infty}\alpha_{j-1}^2\\ <&\infty\end{aligned}$$

   as $\sum\alpha_k^2$ converges by Assumption 2.1.

   Finally, $\sum_{k=0}^{\infty}\gamma_k\alpha_k\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\|<\infty$.

D Convergence Rate

As a reminder of the definitions of the parameters in Theorem 2.7, consider again $R=\|\mathbf{x}_0-\mathbf{1}\bar{x}_0\|^2$, $\delta_k=\big(\frac{1+\rho_w^2}{2}\big)^k$, and $\beta_k=\sum_{j=1}^{k}\delta_j\alpha_{k-j}^2$. From Lemma A.1 and (28), we let $\bar{M}$ denote the upper bound of $\mathbb{E}[\|\bar{g}_k\|^2]$, and from (36), G that of $\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\|$. Let again $\bar{G}=\frac{2\rho_w^2}{1-\rho_w^2}G^2$.

Our primary result, stated in the following Lemma, is based on finding a relation between two successive iterations of the expected divergence.

Lemma D.1. *Let $A=\lambda\alpha_2$, $B=\frac{2\alpha_2L^2}{\lambda n}$, and $C=\frac{\alpha_1^2\alpha_3^6}{2\alpha_2\lambda}$. Whenever Assumption 2.6 holds, for $k>1$, we get*

$$D_{k+1}\leq(1-A\alpha_k\gamma_k)D_k+B\alpha_k\gamma_k[\delta_kR+\beta_k\bar{G}]+\bar{M}\alpha_k^2+C\alpha_k\gamma_k^3. \tag{38}$$

Proof: See D.2.

Next, we let

$$K_0 = \arg \min_{A\alpha_k\gamma_k<1} k.$$

For the ensuing part, the purpose is to locate a vanishing upper bound of D_k, making use of the inequality (38). The idea is to propose a decreasing sequence $U_{k+1} \leq U_k$ and suppose that $D_k \leq U_k$, $\forall k \geq K_0$, and then verify that $D_{k+1} \leq U_{k+1}$ by induction. The choice of U_k is the most difficult component as one has to keep in mind the general forms of α_k and γ_k in (38) and what kind of decisions to take regarding these forms, alongside knowing the exact rate of β_k. An essential property of U_k is presented in the subsequent lemma.

Lemma D.2. *If a decreasing sequence $U_{k+1} \leq U_k$ for $k \geq K_0$ exists such that $D_{k+1} \leq U_{k+1}$ can be deduced from $D_k \leq U_k$ and (38), then*

$$U_k \geq \frac{B}{A}[\delta_k R + \beta_k \bar{G}] + \frac{\bar{M}}{A}\frac{\alpha_k}{\gamma_k} + \frac{C}{A}\gamma_k^2. \tag{39}$$

Proof: See D.3.

An important remark is that the lower bound of U_k in (39) is vanishing as $\delta_k, \beta_k, \frac{\alpha_k}{\gamma_k}$, and γ_k are all vanishing. This lower bound provides an insight on the convergence rate of D_k as it cannot be better than that of $\delta_k, \beta_k, \frac{\alpha_k}{\gamma_k}$, or γ_k^2.

The previous Lemma allows us to move forward in confirming the existence of the constants ς_1 and ς_2 that permit $D_k \leq \varsigma_1\gamma_k^2$ and $D_k \leq \varsigma_2\frac{\alpha_k}{\gamma_k}$ in Theorem 2.7, respectively.

D.1 Proof of Theorem 2.7

1. **Proof of** (12)

   By definition of ς_1, $D_{K_0} \leq \varsigma_1\gamma_{K_0}^2$. The next step is to make sure that $D_{k+1} \leq U_{k+1}$ can be obtained from $D_k \leq U_k$, $\forall k \geq K_0$. Take $U_k = \varsigma_1\gamma_k^2$, let $D_k \leq U_k$ hold, and substitute in (38),

   $$D_{k+1} \leq (1 - A\alpha_k\gamma_k)\varsigma_1\gamma_k^2 + B\alpha_k\gamma_k[\delta_k R + \beta_k\bar{G}] + \bar{M}\alpha_k^2 + C\alpha_k\gamma_k^3.$$

   We solve $D_{k+1} \leq U_{k+1}$ for $\varsigma_1 \in \mathbb{R}^+$

   $$(1 - A\alpha_k\gamma_k)\varsigma_1\gamma_k^2 + B\alpha_k\gamma_k[\delta_k R + \beta_k\bar{G}] + \bar{M}\alpha_k^2 + C\alpha_k\gamma_k^3 \leq U_{k+1} = \varsigma_1\gamma_{k+1}^2.$$

   Then, by considering $\kappa_k = \frac{1-(\frac{\gamma_{k+1}}{\gamma_k})^2}{\alpha_k\gamma_k} > 0$ as given in (11),

   $$B[\delta_k R + \beta_k\bar{G}]\gamma_k^{-2} + \bar{M}\alpha_k\gamma_k^{-3} + C \leq \varsigma_1(A - \kappa_k),$$

   and assuming $A - \kappa_k > 0$, we find a constant $\bar{\varsigma_1}$ such that

   $$\varsigma_1 \geq \bar{\varsigma_1} = \frac{BR\delta_k\gamma_k^{-2} + B\bar{G}\beta_k\gamma_k^{-2} + \bar{M}\alpha_k\gamma_k^{-3} + C}{A - \kappa_k},$$

   keeping in mind that $BR\delta_k\gamma_k^{-2} + B\bar{G}\beta_k\gamma_k^{-2} + \bar{M}\alpha_k\gamma_k^{-3} + C$ is positive by definition. Examine the parameters σ_1, σ_2, σ_3 and σ_4 as they are introduced in (11), then

   $$\bar{\varsigma_1} \leq \frac{BR\sigma_2 + B\bar{G}\sigma_3 + \bar{M}\sigma_4 + C}{A - \sigma_1},$$

   We conclude that $D_k \leq \varsigma_1\gamma_k^2$ where ς_1 satisfies the definition (13).

2. **Proof of** (14)

   $D_{K_0} \leq \varsigma_2\frac{\gamma_{K_0}}{\alpha_{K_0}}$ by definition of ς_2. $\forall k \geq K_0$, let $D_k \leq \varsigma_2\frac{\alpha_k}{\gamma_k}$, then

   $$D_{k+1} \leq (1 - A\alpha_k\gamma_k)\varsigma_2\frac{\alpha_k}{\gamma_k} + B\alpha_k\gamma_k[\delta_k R + \beta_k\bar{G}] + \bar{M}\alpha_k^2 + C\alpha_k\gamma_k^3.$$

Solving $D_{k+1} \leq \varsigma_2 \frac{\alpha_{k+1}}{\gamma_{k+1}}$ for $\varsigma_2 \in \mathbb{R}^+$,

$$(1 - A\alpha_k\gamma_k)\varsigma_2 \frac{\alpha_k}{\gamma_k} + B\alpha_k\gamma_k[\delta_k R + \beta_k \bar{G}] + \bar{M}\alpha_k^2 + C\alpha_k\gamma_k^3 \leq \varsigma_2 \frac{\alpha_{k+1}}{\gamma_{k+1}}.$$

Take $\tau_k = \frac{\frac{\alpha_k}{\gamma_k} - \frac{\alpha_{k+1}}{\gamma_{k+1}}}{\alpha_k^2} > 0$ as given in (11), then

$$B\gamma_k\alpha_k^{-1}[\delta_k R + \beta_k \bar{G}] + C\gamma_k^3\alpha_k^{-2} + \bar{M} \leq (A - \tau_k)\varsigma_2.$$

If $\frac{\alpha_k}{\gamma_k} - \frac{\alpha_{k+1}}{\gamma_{k+1}} < A\alpha_k^2$, then $\exists\ \bar{\varsigma_2}$ such that

$$\varsigma_2 \geq \bar{\varsigma_2} = \frac{BR\delta_k\gamma_k\alpha_k^{-1} + B\bar{G}\beta_k\gamma_k\alpha_k^{-1} + C\gamma_k^3\alpha_k^{-1} + \bar{M}}{(A - \tau_k)}.$$

Examine $\sigma_5, \sigma_6, \sigma_7$ and σ_8 that are defined in (11), we can say

$$\bar{\varsigma_2} \leq \frac{BR\sigma_7 + B\bar{G}\sigma_8 + C\sigma_6 + \bar{M}}{(A - \sigma_5)}.$$

We conclude that $D_k \leq \varsigma_2 \frac{\alpha_k}{\gamma_k}$ with ς_2 satisfying (15).

D.2 Proof of Lemma D.1

We start by expressing the expected divergence in terms of its previous iteration.

$$\begin{aligned}
D_{k+1} =& \mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2] \\
=& \mathbb{E}[\|\bar{x}_k - \alpha_k\bar{g}_k - x^*\|^2] \\
=& D_k + \alpha_k^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha_k\mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k\rangle] \\
\overset{(a)}{=}& D_k + \alpha_k^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, h(\mathbf{x}_k) + \bar{b}_k\rangle] \\
=& D_k + \alpha_k^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k)\rangle] + 2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle] \\
&- 2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \bar{b}_k\rangle]
\end{aligned} \tag{40}$$

where (a) is due to both $\mathbb{E}[e_k|\mathcal{H}_k] = 0$ and (21):

$$\begin{aligned}
\mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k\rangle] &= \mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k] + \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= \mathbb{E}[\langle \bar{x}_k - x^*, e_k\rangle] + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= \mathbb{E}_{\mathcal{H}_k}[\mathbb{E}[\langle \bar{x}_k - x^*, e_k\rangle|\mathcal{H}_k]] + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= 0 + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle].
\end{aligned}$$

From Lemma A.1 and (28), we have $\mathbb{E}[\|\bar{g}_k\|^2] < \bar{M}$ almost surely, with $\bar{M} = \frac{1}{n}M$ a bounded constant. By the strong convexity in Assumption 2.6, we have

$$\begin{aligned}
-2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k)\rangle] &\leq -2\lambda\alpha_2\alpha_k\gamma_k\mathbb{E}[\|\bar{x}_k - x^*\|^2] \\
&= -2\lambda\alpha_2\alpha_k\gamma_k D_k.
\end{aligned} \tag{41}$$

Next, from Lemma 1.6, we have

$$\begin{aligned}
2\alpha_2\alpha_k\gamma_k\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle &\leq 2\alpha_2\alpha_k\gamma_k\frac{L}{\sqrt{n}}\|\bar{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \\
&\overset{(a)}{\leq} \frac{\lambda\alpha_2\alpha_k\gamma_k}{2}\|\bar{x}_k - x^*\|^2 + 2\alpha_2\alpha_k\gamma_k\frac{L^2}{\lambda n}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2,
\end{aligned}$$

where (a) is due to $2\sqrt{\epsilon} \times \frac{1}{\sqrt{\epsilon}}\langle a, b\rangle = 2\langle \sqrt{\epsilon}a, \frac{1}{\sqrt{\epsilon}}b\rangle \leq \epsilon\|a\|^2 + \frac{1}{\epsilon}\|b\|^2$. From (34) and (36), we get

$$\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \leq \delta_k R + \beta_k \bar{G}$$

where $R = \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2$, $\delta_k = \left(\frac{1+\rho_w^2}{2}\right)^k$, $\beta_k = \sum_{j=1}^k \delta_j \alpha_{k-j}^2$, and $\bar{G} = \frac{2\rho_w^2}{1-\rho_w^2}G^2$. Hence,

$$2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle] \leq \frac{\lambda\alpha_2\alpha_k\gamma_k}{2}D_k + 2\alpha_2\alpha_k\gamma_k\frac{L^2}{\lambda n}[\delta_k R + \beta_k\bar{G}]. \tag{42}$$

From (22),

$$\begin{aligned} -2\alpha_2\alpha_k\gamma_k\mathbb{E}[\langle \bar{x}_k - x^*, \bar{b}_k\rangle] &\leq \frac{\lambda\alpha_2\alpha_k\gamma_k}{2}D_k + \frac{2\alpha_2\alpha_k\gamma_k}{\lambda}\mathbb{E}[\|\bar{b}_k\|^2] \\ &\leq \frac{\lambda\alpha_2\alpha_k\gamma_k}{2}D_k + \frac{\alpha_1^2\alpha_3^6\alpha_k\gamma_k^3}{2\alpha_2\lambda} \end{aligned} \tag{43}$$

Finally, by combining (40), (41), (42), and (43), we get (38).

D.3 Proof of Lemma D.2

Since $1 - A\alpha_k\gamma_k > 0$ when $k \geq K_0$, we may substitute $D_k \leq U_k$ in (38),

$$D_{k+1} \leq (1 - \lambda\alpha_2\alpha_k\gamma_k)U_k + 2\alpha_2\alpha_k\gamma_k\frac{L^2}{\lambda n}[\delta_k R + \beta_k\bar{G}] + \alpha_k^2\bar{M} + \frac{\alpha_1^2\alpha_3^6\alpha_k\gamma_k^3}{2\alpha_2\lambda}.$$

Testing $D_{k+1} \leq U_{k+1}$ in the previous inequality, we get

$$(1 - \lambda\alpha_2\alpha_k\gamma_k)U_k + 2\alpha_2\alpha_k\gamma_k\frac{L^2}{\lambda n}[\delta_k R + \beta_k\bar{G}] + \alpha_k^2\bar{M} + \frac{\alpha_1^2\alpha_3^6\alpha_k\gamma_k^3}{2\alpha_2\lambda} \leq U_{k+1} \leq U_k$$

$$\frac{2L^2}{\lambda^2 n}[\delta_k R + \beta_k\bar{G}] + \frac{\bar{M}}{\lambda\alpha_2}\frac{\alpha_k}{\gamma_k} + \frac{\alpha_1^2\alpha_3^6}{2\lambda^2\alpha_2^2}\gamma_k^2 \leq U_k. \tag{44}$$

D.4 Proof of Lemma 2.8

We consider the following series, as it affects the maximum possible convergence rate which appears in (39) and all inequalities related to D_k,

$$\beta_k = \sum_{j=1}^k \delta_j\alpha_{k-j}^2 = \sum_{j=1}^k \left(\frac{1+\rho_w^2}{2}\right)^j \alpha_{k-j}^2.$$

We then try to find β_{k+1} in terms of β_k,

$$\begin{aligned} \beta_{k+1} &= \sum_{j=1}^{k+1}\left(\frac{1+\rho_w^2}{2}\right)^j \alpha_{k+1-j}^2 \\ &= \sum_{j=2}^{k+1}\left(\frac{1+\rho_w^2}{2}\right)^j \alpha_{k+1-j}^2 + \left(\frac{1+\rho_w^2}{2}\right)\alpha_k^2 \\ &= \sum_{j=1}^{k}\left(\frac{1+\rho_w^2}{2}\right)^{j+1} \alpha_{k-j}^2 + \left(\frac{1+\rho_w^2}{2}\right)\alpha_k^2 \\ &= \left(\frac{1+\rho_w^2}{2}\right)(\beta_k + \alpha_k^2). \end{aligned}$$

We know that this series has a Q-linear to a Q-sublinear convergence rate as

$$\begin{aligned}
\frac{\beta_{k+1}}{\beta_k} &= \left(\frac{1+\rho_w^2}{2}\right)\left(1+\frac{\alpha_k^2}{\sum_{j=1}^{k}\left(\frac{1+\rho_w^2}{2}\right)^j\alpha_{k-j}^2}\right)\\
&= \left(\frac{1+\rho_w^2}{2}\right)+\frac{\alpha_k^2}{\sum_{j=1}^{k}\left(\frac{1+\rho_w^2}{2}\right)^{j-1}\alpha_{k-j}^2}\\
&= \left(\frac{1+\rho_w^2}{2}\right)+\frac{\alpha_k^2}{\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j}\alpha_{k-1-j}^2}\\
&= \left(\frac{1+\rho_w^2}{2}\right)+\frac{1}{\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j}\frac{\alpha_{k-1-j}^2}{\alpha_k^2}}\\
&\overset{(a)}{\leq} \left(\frac{1+\rho_w^2}{2}\right)+\frac{1}{\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^{j}}\\
&\overset{(b)}{=} \left(\frac{1+\rho_w^2}{2}\right)+\frac{1-\left(\frac{1+\rho_w^2}{2}\right)}{1-\left(\frac{1+\rho_w^2}{2}\right)^{k}}
\end{aligned}$$

where (a) is since $\frac{\alpha_{k-1-j}^2}{\alpha_k^2} \geq 1$ for every $j \in \{0,\ldots,k-1\}$ and (b) is due to the geometric sum $\sum_{j=0}^{k-1}\left(\frac{1+\rho_w^2}{2}\right)^j = \frac{1-\left(\frac{1+\rho_w^2}{2}\right)^k}{1-\left(\frac{1+\rho_w^2}{2}\right)}$. Hence,

$$\lim_{k\to\infty}\frac{\beta_{k+1}}{\beta_k} \leq \left(\frac{1+\rho_w^2}{2}\right)+1-\left(\frac{1+\rho_w^2}{2}\right)=1$$

since $\lim_{k\to\infty}\left(\frac{1+\rho_w^2}{2}\right)^k = 0$. Next, to get an idea of how β_k converges in terms of k, we consider K_1 that is defined in (17). We know that K_1 exists and is finite as $\left(\frac{1+\rho_w^2}{2}\right)^k$ decreases much faster than α_k^2. Taking α_k as that in (16), we know that $0.5 < \upsilon_1 < 1$ and we find the condition on υ_1 such that

$$\begin{aligned}
\left(\frac{1+\rho_w^2}{2}\right)^k &\leq \alpha_k^2\\
k\log\left(\frac{1+\rho_w^2}{2}\right) &\leq -2\upsilon_1\log(k+1)\\
\upsilon_1 &\leq \frac{1}{2}\log\left(\frac{2}{1+\rho_w^2}\right)\frac{k}{\log(k+1)},
\end{aligned}$$

which is feasible as $\frac{1}{2}\log\left(\frac{2}{1+\rho_w^2}\right) > 0$ and $\frac{k}{\log(k+1)}$ grows very large for $k \geq K_1$, a simple condition is that $1 < \frac{1}{2}\log(\frac{2}{1+\rho_w^2})\frac{k}{\log(k+1)}$ which gives

$$K_1 = \arg\min_{2\log^{-1}(\frac{2}{1+\rho_w^2})<k\log^{-1}(k+1)} k.$$

Thus, we can write β_k as

$$\begin{aligned}
\beta_k &= \sum_{j=1}^{K_1-1} \left(\frac{1+\rho_w^2}{2}\right)^j \alpha_{k-j}^2 + \sum_{j=K_1}^{k} \left(\frac{1+\rho_w^2}{2}\right)^j \alpha_{k-j}^2 \\
&\overset{(a)}{\leq} \alpha_{k-K_1}^2 \sum_{j=1}^{K_1-1} \left(\frac{1+\rho_w^2}{2}\right)^j + \sum_{j=K_1}^{k} \alpha_j^2 \alpha_{k-j}^2 \\
&\overset{(b)}{=} \alpha_{k-K_1}^2 \left(\frac{1+\rho_w^2}{1-\rho_w^2}\right)\left(1 - \left(\frac{1+\rho_w^2}{2}\right)^{K_1-1}\right) + \sum_{j=K_1}^{k} \alpha_j^2 \alpha_{k-j}^2 \\
&= \beta' \alpha_{k-K_1}^2 + \sum_{j=K_1}^{k} \alpha_j^2 \alpha_{k-j}^2,
\end{aligned} \tag{45}$$

where in (a) we used the definition of K_1 and the fact that α_k^2 is a decreasing step-size, and in (b) the sum of a geometric series. We know that $\beta' = \left(\frac{1+\rho_w^2}{1-\rho_w^2}\right)\left(1 - \left(\frac{1+\rho_w^2}{2}\right)^{K_1-1}\right)$ is finite, then the first sum decreases as

$\alpha_{k-K_1}^2 = \frac{\alpha_0^2}{(k-K_1+1)^{2\upsilon_1}} \sim \frac{1}{k^{2\upsilon_1}}$ with $1 < 2\upsilon_1 < 2$, which leaves us with the second sum,

$$
\begin{aligned}
\sum_{j=K_1}^{k} \alpha_j^2 \alpha_{k-j}^2 &= \alpha_0^2 \sum_{j=K_1}^{k} \frac{1}{(j+1)^{2\upsilon_1}} \frac{1}{(k-j+1)^{2\upsilon_1}} \\
&= \alpha_0^2 \sum_{j=K_1+1}^{k+1} \frac{1}{j^{2\upsilon_1}} \frac{1}{(k-j+2)^{2\upsilon_1}} \\
&= \alpha_0^2 \sum_{j=K_1+1}^{k+1} \frac{1}{j^{2\upsilon_1}} \frac{1}{(k+2)^{2\upsilon_1}(1-\frac{j}{k+2})^{2\upsilon_1}} \\
&\overset{(a)}{=} \alpha_0^2 \frac{1}{(k+2)^{4\upsilon_1}} \sum_{u \in \mathcal{U}_k} \Big(\frac{1}{u(1-u)}\Big)^{2\upsilon_1} \\
&= \alpha_0^2 \frac{|\mathcal{U}_k|}{(k+2)^{4\upsilon_1}} \sum_{u \in \mathcal{U}_k} \Big(\frac{1}{u^2(1-u)^2}\Big)^{\upsilon_1} \times \frac{1}{|\mathcal{U}_k|} \\
&\overset{(b)}{\leq} \alpha_0^2 \frac{|\mathcal{U}_k|}{(k+2)^{4\upsilon_1}} \Bigg(\sum_{u \in \mathcal{U}_k} \frac{1}{u^2(1-u)^2} \times \frac{1}{|\mathcal{U}_k|} \Bigg)^{\upsilon_1} \\
&\overset{(c)}{=} \alpha_0^2 \frac{|\mathcal{U}_k|^{1-\upsilon_1}}{(k+2)^{4\upsilon_1}} \Bigg(\sum_{u \in \mathcal{U}_k} \frac{1}{u^2} + \frac{1}{(1-u)^2} + \frac{2}{u} + \frac{2}{(1-u)} \Bigg)^{\upsilon_1} \\
&= \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{4\upsilon_1}} \Bigg(\sum_{u \in \mathcal{U}_k} \frac{1}{u^2} + \frac{2}{u} + \sum_{u \in \mathcal{U}_k \setminus \{\frac{k+1}{k+2}\}} \frac{1}{(1-u)^2} + \frac{2}{(1-u)} \\
&\qquad\qquad + 2(k+2) + (k+2)^2 \Bigg)^{\upsilon_1} \qquad (46) \\
&\overset{(d)}{\leq} \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{4\upsilon_1}} \Bigg((k+2) \int_{\frac{K_1}{k+2}}^{\frac{k+1}{k+2}} \Big(\frac{1}{u^2} + \frac{2}{u}\Big) du \\
&\qquad + (k+2) \int_{\frac{K_1+1}{k+2}}^{\frac{k+1}{k+2}} \Big(\frac{1}{(1-u)^2} + \frac{2}{(1-u)}\Big) du + 2(k+2) + (k+2)^2 \Bigg)^{\upsilon_1} \\
&= \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{3\upsilon_1}} \Bigg(\frac{(k+2)(k-K_1+1)}{K_1(k+1)} \\
&\qquad + 2\ln \frac{e(k+1)(k-K_1+1)}{K_1} + \frac{(k+2)(k-K_1)}{k-K_1+1} + (k+2) \Bigg)^{\upsilon_1} \\
&\leq \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{3\upsilon_1}} \Bigg(\frac{(k+2)(k+1)}{K_1(k+1)} \\
&\qquad + 2\ln \frac{e(k+1)(k-K_1+1)}{K_1} + \frac{(k+2)(k-K_1)}{k-K_1} + (k+2) \Bigg)^{\upsilon_1} \\
&= \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{3\upsilon_1}} \Bigg(2\ln \frac{e(k+1)(k-K_1+1)}{K_1} + (k+2)\Big(2 + \frac{1}{K_1}\Big) \Bigg)^{\upsilon_1} \\
&\leq \alpha_0^2 \frac{(k-K_1+1)^{1-\upsilon_1}}{(k+2)^{3\upsilon_1}} \Bigg(4\ln \frac{e(k+2)}{K_1} + (k+2)\Big(2 + \frac{1}{K_1}\Big) \Bigg)^{\upsilon_1} \\
&\leq \alpha_0^2 \frac{(k+2)^{1-\upsilon_1}}{(k+2)^{3\upsilon_1}} (k+2)^{\upsilon_1} \Bigg(\frac{4}{(k+2)} \ln \frac{e}{K_1} + 6 + \frac{1}{K_1} \Bigg)^{\upsilon_1} \\
&\leq \beta'' \frac{1}{(k+2)^{3\upsilon_1 - 1}},
\end{aligned}
$$

where in (a) we changed the summation variable to $u = \frac{j}{k+2}$ and $\mathcal{U}_k = \{\frac{K_1+1}{k+2}, \frac{K_1+2}{k+2}, \ldots, \frac{k+1}{k+2}\}$, in (b) we used Jensen's inequality $\mathbb{E}[\varphi(x)] \leq \varphi(\mathbb{E}[x])$ for the concave function $\varphi(x) = x^{\upsilon_1}$ with $0.5 < \upsilon_1 < 1$, and (c)

is by partial fraction decomposition. In (d), we interpret the sum over $\mathcal{U}_k$ as a Riemann sum in which the function $\frac{1}{u}$ is evaluated at the right endpoint of the interval $[\frac{K_1+i}{k+2}, \frac{K_1+i+1}{k+2}]$, for $i = 0, 1, \ldots, k - K_1$. Since the function $\frac{1}{u}$ is monotonically decreasing, it is in fact a *lower* Riemann sum and therefore bounded from above by the integral $\int_{\frac{K_1}{k+2}}^{\frac{k+1}{k+2}} \frac{1}{u} du$. Analogously, we estimate the sum of $\frac{1}{u^2}$ over $\mathcal{U}_k$. Mutatis mutandis, the estimate for the monotonically increasing functions $\frac{1}{1-u}$ and $\frac{1}{(1-u)^2}$ follows.

We have also let $\beta'' = \max_{k \geq K_1} \left(\frac{4}{(k+2)} \ln \frac{e}{K_1} + 6 + \frac{1}{K_1} \right)^{\upsilon_1} = \left(\frac{4}{(K_1+2)} \ln \frac{e}{K_1} + 6 + \frac{1}{K_1} \right)^{\upsilon_1}$. Hence, $\beta'' \frac{1}{(k+2)^{3\upsilon_1 - 1}} \sim \frac{1}{k^{3\upsilon_1 - 1}}$ with $0.5 < 3\upsilon_1 - 1 < 2$.

Since $3\upsilon_1 - 1 < 2\upsilon_1$ for $0.5 < \upsilon_1 < 1$, then $\frac{1}{k^{2\upsilon_1}} < \frac{1}{k^{3\upsilon_1 - 1}}$ for all $k \geq K_1$.

From (45) and (46),

$$
\begin{aligned}
\beta_k &\leq \beta' \frac{1}{(k - K_1 + 1)^{2\upsilon_1}} + \beta'' \frac{1}{(k+2)^{3\upsilon_1 - 1}} \\
&\leq \beta' \frac{1}{(k - K_1 + 1)^{3\upsilon_1 - 1}} + \beta'' \frac{1}{(k - K_1 + 1)^{3\upsilon_1 - 1}} \\
&= (\beta' + \beta'') \frac{1}{(k - K_1 + 1)^{3\upsilon_1 - 1}},
\end{aligned}
$$

and we deduce that β_k has a convergence rate of at least $\frac{1}{k^{3\upsilon_1 - 1}}$, which concludes the proof.

D.5 Proof of Theorem 2.9

Theorem 2.7 indicates that the convergence rate is a function of υ_1 and υ_2, as $\gamma_k^2 \propto (k+1)^{-2\upsilon_2}$ and $\frac{\alpha_k}{\gamma_k} \propto (k+1)^{-(\upsilon_1 - \upsilon_2)}$. Nonetheless, we must still verify the validity of the assumptions presented in the theorem, meaning:

- Are $\sigma_1 < A$ and $\sigma_5 < A$ fulfilled?
- Are ς_1 and ς_2 bounded?

We must remark that in what follows, the analysis is done for $k \geq K_2$.

Let α_k and γ_k have the forms given in (16).

1. **Verifying that $\sigma_1 < A$ and $\sigma_5 < A$**

   The idea is to find a bound on α_0 and γ_0 to guarantee $\sigma_1 < A$ and $\sigma_5 < A$. We start by bounding σ_1 and σ_5 from above, i.e.,

$$
\sigma_1 = \max_{k \geq K_2} \frac{1 - (\frac{\gamma_{k+1}}{\gamma_k})^2}{\alpha_k \gamma_k} = \max_{k \geq K_2} \frac{1 - (1 + \frac{1}{k+1})^{-2\upsilon_2}}{\alpha_0 \gamma_0 (k+1)^{-\upsilon_1 - \upsilon_2}}
$$

   and

$$
\sigma_5 = \max_{k \geq K_2} \frac{1 - \frac{\alpha_{k+1} \gamma_{k+1}^{-1}}{\alpha_k \gamma_k^{-1}}}{\alpha_k \gamma_k} = \max_{k \geq K_2} \frac{1 - (1 + \frac{1}{k+1})^{-(\upsilon_1 - \upsilon_2)}}{\alpha_0 \gamma_0 (k+1)^{-\upsilon_1 - \upsilon_2}}.
$$

   To do so, we define a function $q(x) = x^{-a}(1 - (1+x)^{-b})$ with $a, b, x \in (0, 1]$. Since $x^{-a} \leq x^{-1}$, we have $q(x) \leq x^{-1}(1 - (1+x)^{-b}) = r(x)$. To further bound $q(x)$, We study the derivative of $r(x)$ as it is simpler to do so,

$$
r'(x) = x^{-2} \Big(((b+1)x + 1)(1+x)^{-b-1} - 1 \Big) = x^{-2} s(x).
$$

Hence the sign of $r'(x)$ is that of $s(x)$. We again calculate the derivative of $s(x)$ to find its sign,

$$s'(x) = -b(b+1)x(1+x)^{-b-2} \leq 0$$

since $b > 0$ and $x > 0$. Then, $s(x)$ is a decreasing function of x over $(0, 1]$. We remark that $\lim_{x\to 0} s(x) = 0$, meaning $s(x) < 0$ and $r'(x) < 0$, $\forall x \in (0, 1]$. Finally,

$$r(x) < \lim_{x\to 0} r(x) = \frac{1-(1+x)^{-b}}{x} = b,$$

and $q(x) \leq r(x) < b$, noting that $\lim_{x\to 0} q(x) = b$ for $a = 1$. We conclude that $\sigma_1 < \frac{2\upsilon_2}{\alpha_0\gamma_0}$ and $\sigma_1 < \frac{\upsilon_1-\upsilon_2}{\alpha_0\gamma_0}$. For $\sigma_1 < A$ and $\sigma_5 < A$ to be valid, we must have

$$\alpha_0\gamma_0 \geq \max\{2\upsilon_2, \upsilon_1 - \upsilon_2\}/A. \tag{47}$$

2. **Verifying that ς_1 and ς_2 are bounded**

The goal is to verify that the constant term in the convergence rate is bounded. Thus, we must check that the lower bounds given in (13) and (15) are indeed finite. We begin by analyzing σ_2 and σ_7, i.e.,

$$\sigma_2 = \max_{k\geq K_2} \frac{\delta_k}{\gamma_k^2} = \max_{k\geq K_2} \gamma_0^{-2}(k+1)^{2\upsilon_2}\left(\frac{1+\rho_w^2}{2}\right)^k$$

and

$$\sigma_7 = \max_{k\geq K_2} \frac{\gamma_k}{\alpha_k}\delta_k = \max_{k\geq K_2} \gamma_0\alpha_0^{-1}(k+1)^{\upsilon_1-\upsilon_2}\left(\frac{1+\rho_w^2}{2}\right)^k.$$

To prove that σ_2 and σ_7 are finite, we define the function $p(k) = a(k+1)^b\left(\frac{1+\rho_w^2}{2}\right)^k$, with $a > 0$ and $0 < b < 1$. To find the maximum, we find k such that $p'(k) = 0$, meaning

$$a(k+1)^b\left(\frac{1+\rho_w^2}{2}\right)^k\left(\frac{b}{k+1} + \ln\left(\frac{1+\rho_w^2}{2}\right)\right) = 0.$$

Hence,

$$k = b\ln^{-1}\left(\frac{2}{1+\rho_w^2}\right) - 1.$$

- From here, we can say if $b\ln^{-1}\left(\frac{2}{1+\rho_w^2}\right) - 1 \geq K_2$, then

$$\max_{k\geq K_2} p(k) = \frac{ab^b}{\left(\frac{2}{1+\rho_w^2}\right)\ln^b\left(\frac{2}{1+\rho_w^2}\right)}e^{-b} < \infty.$$

- Otherwise, if $b\ln^{-1}\left(\frac{2}{1+\rho_w^2}\right) - 1 < K_2 \leq k$, then $\frac{b}{k+1} + \ln\left(\frac{1+\rho_w^2}{2}\right) < 0$ which gives $p' < 0$ for $k \geq K_2$, meaning p is strictly decreasing and

$$\max_{k\geq K_2} p(k) = p(K_2) = a(K_2+1)^b\left(\frac{1+\rho_w^2}{2}\right)^{K_2} < \infty.$$

We conclude that $\sigma_2 < \infty$ and $\sigma_7 < \infty$.

Next, we study the finiteness of σ_3 and σ_8. From Lemma 2.8, we can write

$$\begin{aligned}\sigma_3 = \max_{k\geq K_2} \frac{\beta_k}{\gamma_k^2} &\leq \max_{k\geq K_2} (\beta' + \beta'')\frac{\alpha_{k-K_1}^2}{\gamma_k^2} \\ &= \max_{k\geq K_2} (\beta' + \beta'')\alpha_0^2\gamma_0^{-2}\frac{(k+1)^{2\upsilon_2}}{(k-K_1+1)^{3\upsilon_1-1}}\end{aligned}$$

and

$$\sigma_8 = \max_{k\geq K_2} \frac{\gamma_k}{\alpha_k}\beta_k \leq \max_{k\geq K_2} (\beta' + \beta'')\frac{\gamma_k}{\alpha_k} = \frac{\gamma_k \alpha_{k-K_1}^2}{\alpha_k}$$
$$= \max_{k\geq K_2} (\beta' + \beta'')\gamma_0\alpha_0 \frac{(k+1)^{\upsilon_1-\upsilon_2}}{(k-K_1+1)^{3\upsilon_1-1}}.$$

We then define a function $q(k) = a\frac{(k+1)^b}{(k-K_1+1)^{3\upsilon_1-1}}$ for $k \geq K_2$, with $a > 0$ and $0 < b < 1$, and we study its derivative

$$q'(k) = a\frac{(b-3\upsilon_1+1)k - bK_1 + b - 3\upsilon_1 + 1}{(k+1)^{1-b}(k-K_1+1)^{3\upsilon_1}}.$$

We know that $q' < 0$, and thus q is strictly decreasing for $k \geq K_2$, when $b - 3\upsilon_1 + 1 \leq 0$.

Hence,

$$\sigma_3 = \begin{cases} \alpha_0^2\gamma_0^{-2}\frac{(K_2+1)^{2\upsilon_2}}{(K_2-K_1+1)^{3\upsilon_1-1}}, & \text{if } 2\upsilon_2 - 3\upsilon_1 + 1 \leq 0, \\ \infty, & \text{if } 2\upsilon_2 - 3\upsilon_1 + 1 > 0, \end{cases} \tag{48}$$

and

$$\sigma_8 = (\beta' + \beta'')\gamma_0\alpha_0 \frac{(K_2+1)^{\upsilon_1-\upsilon_2}}{(K_2-K_1+1)^{3\upsilon_1-1}} < \infty$$

since $b - 3\upsilon_1 + 1 = -\upsilon_2 - 2\upsilon_1 + 1 < 0$ always holds.

We end with the analysis of σ_4 and σ_6, i.e.,

$$\sigma_4 = \alpha_0\gamma_0^{-3}\max_{k\geq K_2} (1+k)^{-(\upsilon_1-3\upsilon_2)} = \begin{cases} \alpha_0\gamma_0^{-3}(1+K_2)^{-(\upsilon_1-3\upsilon_2)}, & \text{if } \upsilon_1 \geq 3\upsilon_2, \\ \infty, & \text{if } \upsilon_1 < 3\upsilon_2, \end{cases}$$

and

$$\sigma_6 = \alpha_0^{-1}\gamma_0^{3}\max_{k\geq K_2} (1+k)^{\upsilon_1-3\upsilon_2} = \begin{cases} \alpha_0^{-1}\gamma_0^{3}(1+K_2)^{\upsilon_1-3\upsilon_2}, & \text{if } \upsilon_1 \leq 3\upsilon_2, \\ \infty, & \text{if } \upsilon_1 > 3\upsilon_2. \end{cases}$$

There are clearly 3 cases:

- $\upsilon_1 > 3\upsilon_2$
  Thus, σ_4 is bounded.
  Since now $2\upsilon_2 - 3\upsilon_1 + 1 < \frac{2}{3}\upsilon_1 - 3\upsilon_1 + 1 = -\frac{7}{3}\upsilon_1 + 1 < 0$ always holds, then σ_3 (48) and ς_1 (by definition) are also bounded provided that $\alpha_0\gamma_0 \geq \frac{2\upsilon_2}{A}$ in (47).
  However, $\varsigma_2 \to \infty$ since $\sigma_6 \to \infty$ resulting in a loose upper bound in (14).
  To that end, we can write $D_k \leq \Upsilon_1(1+k)^{-2\upsilon_2}$ with Υ_1 a bounded constant.
- $\upsilon_1 < 3\upsilon_2$
  Similarly, σ_6 is bounded while $\sigma_4 \to \infty$. Then, $\exists\ \Upsilon_2 < \infty$, where $D_k \leq \Upsilon_2(1+k)^{-(\upsilon_1-\upsilon_2)}$ provided that $\alpha_0\gamma_0 \geq \frac{\upsilon_1-\upsilon_2}{A}$.
- $\upsilon_1 = 3\upsilon_2$
  Both σ_4 and σ_6 are bounded allowing both previous inequalities corresponding to D_k to be valid.

By this analysis, we conclude the proof of Theorem 2.9.

E Regret Analysis

Since, by Lemma 1.6, the objective function is L-smooth, we can write

$$\mathcal{F}(y) \leq \mathcal{F}(x) + \langle \nabla\mathcal{F}(x), y - x\rangle + \frac{L}{2}\|y-x\|^2,\ \forall x, y \in \mathbb{R}^d. \tag{49}$$

To find the regret bound, consider

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_k}\left[\sum_{k=1}^{K}\mathcal{F}(\bar{x}_k)-\mathcal{F}(x^*)\right] &\overset{(a)}{\leq} \frac{L}{2}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{H}_k}\left[\|\bar{x}_k-x^*\|^2\right] \\
&= \frac{L}{2}\sum_{k=1}^{K}D_k \\
&\overset{(b)}{\leq} \Upsilon\frac{L}{2}\sum_{k=1}^{K}\frac{1}{\sqrt{k+1}} \\
&\overset{(c)}{\leq} \Upsilon\frac{L}{2}\int_0^K\frac{1}{\sqrt{u+1}}du \\
&= \Upsilon L(\sqrt{K+1}-1)
\end{aligned}$$

where (a) is due to $\nabla\mathcal{F}(x^*)=0$, by definition of x^*, in (49), (b) is by Theorem 2.9, and in (c) we perform the same interpretation of the monotonically decreasing function $\frac{1}{\sqrt{u+1}}$ as in (46).

F Convergence Rate with Constant Step Sizes

We start by going over previous derivations,

$$\begin{aligned}
\breve{g}_{i,k} &= \mathbb{E}_{S,\Phi,\zeta}[\Phi_{i,k}(f_i(x_{i,k}+\gamma\Phi_{i,k},S_{i,k})+\zeta_{i,k})|\mathcal{H}_k] \\
&= \mathbb{E}_{\Phi}[\Phi_{i,k}F_i(x_{i,k}+\gamma\Phi_{i,k})|\mathcal{H}_k] \\
&= F_i(x_{i,k})\mathbb{E}_{\Phi}[\Phi_{i,k}] + \gamma\mathbb{E}_{\Phi}[\Phi_{i,k}\Phi_{i,k}^T|\mathcal{H}_k]\nabla F_i(x_{i,k}) + \frac{\gamma^2}{2}\mathbb{E}_{\Phi}[\Phi_{i,k}\Phi_{i,k}^T\nabla^2F_i(\tilde{x}_{i,k})\Phi_{i,k}|\mathcal{H}_k] \\
&= \alpha_2\gamma[\nabla F_i(x_{i,k})+b_{i,k}].
\end{aligned}$$

Thus, $b_{i,k}=\frac{\gamma}{2\alpha_2}\mathbb{E}_{\Phi}[\Phi_{i,k}\Phi_{i,k}^T\nabla^2F_i(\tilde{x}_{i,k})\Phi_{i,k}|\mathcal{H}_k]$.

Let Assumptions 1.2 and 2.2 hold. Then, we can bound the bias as

$$\begin{aligned}
\|b_{i,k}\| &\leq \frac{\gamma}{2\alpha_2}\mathbb{E}_{\Phi}[\|\Phi_{i,k}\|_2\|\Phi_{i,k}^T\|_2\|\nabla^2F_i(\tilde{x}_{i,k})\|_2\|\Phi_{i,k}\|_2|\mathcal{H}_k] \\
&\leq \gamma\frac{\alpha_3^3\alpha_1}{2\alpha_2}.
\end{aligned}$$

We remark that

$$\begin{aligned}
\tilde{g}_k &= \mathbb{E}[\bar{g}_k|\mathcal{H}_k] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g_{i,k}|\mathcal{H}_k] \\
&= \frac{1}{n}\sum_{i=1}^{n}\alpha_2\gamma[\nabla F_i(x_{i,k})+b_{i,k}] \\
&= \alpha_2\gamma[h(\mathbf{x}_k)+\bar{b}_k]
\end{aligned} \tag{50}$$

is also a biased estimator of $h(\mathbf{x}_k)$ with

$$\begin{aligned}\|\bar{b}_k\| &= \|\frac{1}{n}\sum_{i=1}^{n} b_{i,k}\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\|b_{i,k}\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\gamma\frac{\alpha_3^3\alpha_1}{2\alpha_2} \\ &= \gamma\frac{\alpha_3^3\alpha_1}{2\alpha_2}.\end{aligned} \tag{51}$$

Lemma F.1. *Let all Assumptions 1.3, 1.4, and 2.2 hold and $\|\mathbf{x}_k\| < \infty$ almost surely, then there exists a bounded constant $M > 0$, such that $E[\|\mathbf{g}_k\|^2] < M$ almost surely.*

Proof. $\forall i \in \mathcal{N}$, we have

$$\begin{aligned}\mathbb{E}[\|g_{i,k}\|^2|\mathcal{H}_k] &= \mathbb{E}[\|\Phi_{i,k}(f_i(x_{i,k}+\gamma\Phi_{i,k}, S_{i,k})+\zeta_{i,k})\|^2|\mathcal{H}_k] \\ &= \mathbb{E}[\|\Phi_{i,k}\|^2\|f_i(x_{i,k}+\gamma\Phi_{i,k}, S_{i,k})+\zeta_{i,k}\|^2|\mathcal{H}_k] \\ &\overset{(a)}{\leq} \alpha_3^2\mathbb{E}[(f_i(x_{i,k}+\gamma\Phi_{i,k}, S_{i,k})+\zeta_{i,k})^2|\mathcal{H}_k] \\ &\overset{(b)}{=} \alpha_3^2\mathbb{E}[f_i^2(x_{i,k}+\gamma\Phi_{i,k}, S_{i,k})|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\ &\overset{(c)}{\leq} \alpha_3^2\mathbb{E}[(\|f_i(0, S_{i,k})\| + L_{S_{i,k}}\|x_{i,k}+\gamma\Phi_{i,k}\|)^2|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\ &\overset{(d)}{\leq} 2\alpha_3^2\mathbb{E}[\mu_{S_{i,k}}^2 + L_{S_{i,k}}^2(\|x_{i,k}\|+\gamma\alpha_3)^2|\mathcal{H}_k] + \alpha_3^2\alpha_4 \\ &\overset{(e)}{=} 2\alpha_3^2(\mu + L'(\|x_{i,k}\|+\gamma\alpha_3)^2) + \alpha_3^2\alpha_4 \\ &< \infty,\end{aligned}$$

where (a) is due to Assumption 2.2, (b) Assumption 1.4, and (c) Assumption 1.3. We denote $\|f_i(0, S_{i,k})\| = \mu_{S_{i,k}}$ in (d) and the inequality is due to $\frac{x+y}{2} \leq \sqrt{\frac{x^2+y^2}{2}}$, $\forall x, y \in \mathbb{R}$. In (e), $\mu = \mathbb{E}[\mu_{S_{i,k}}^2]$ and $L' = \mathbb{E}[L_{S_{i,k}}^2]$. □

The stochastic noise is still defined as $e_k = \bar{g}_k - \tilde{g}_k$ and retains its property

$$\mathbb{E}[e_k] = \mathbb{E}[\bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k]] = \mathbb{E}_{\mathcal{H}_k}\Big[\mathbb{E}\Big[\bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\Big|\mathcal{H}_k\Big]\Big] = 0.$$

1. **Proving $\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$ converges linearly**

$$\begin{aligned}&\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \\ =&\|W\mathbf{x}_k - \alpha W\mathbf{y}_k - \mathbf{1}\bar{x}_k + \alpha\mathbf{1}\bar{y}_k\|^2 \\ =&\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 - 2\alpha\langle W\mathbf{x}_k - \mathbf{1}\bar{x}_k, W\mathbf{y}_k - \mathbf{1}\bar{y}_k\rangle + \alpha^2\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\ \overset{(a)}{\leq}&\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \alpha[\frac{1-\rho_w^2}{2\rho_w^2\alpha}\|W\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2\alpha}{1-\rho_w^2}\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \alpha^2\|W\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\ \overset{(b)}{\leq}&\rho_w^2\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \rho_w^2\alpha[\frac{1-\rho_w^2}{2\rho_w^2\alpha}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \frac{2\rho_w^2\alpha}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \rho_w^2\alpha^2\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \\ =&\frac{1+\rho_w^2}{2}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \alpha^2\frac{(1+\rho_w^2)\rho_w^2}{1-\rho_w^2}\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2\end{aligned} \tag{52}$$

where (a) is by $-2\epsilon \times \frac{1}{\epsilon}\langle a, b\rangle = -2\langle \epsilon a, \frac{1}{\epsilon} b\rangle \leq \epsilon^2\|a\|^2 + \frac{1}{\epsilon^2}\|b\|^2$ and (b) is by Lemma 1.5. By induction, we have

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \leq \big(\frac{1+\rho_w^2}{2}\big)^{k+1}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \alpha^2\frac{2\rho_w^2}{1-\rho_w^2}\sum_{j=0}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{j+1}\|\mathbf{y}_{k-j} - \mathbf{1}\bar{y}_{k-j}\|^2. \quad (53)$$

To bound the term $\|\mathbf{y}_{k-j} - \mathbf{1}\bar{y}_{k-j}\|^2$, we repeatedly replace the auxiliary variables with the algorithm's iterations,

$$\begin{aligned}
\mathbf{y}_{k+1} &= W\mathbf{y}_k + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W(W\mathbf{y}_{k-1} + \mathbf{g}_k - \mathbf{g}_{k-1}) + \mathbf{g}_{k+1} - \mathbf{g}_k \\
&= W^2\mathbf{y}_{k-1} - W\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= W^3\mathbf{y}_{k-2} - W^2\mathbf{g}_{k-2} + W(W-I)\mathbf{g}_{k-1} + (W-I)\mathbf{g}_k + \mathbf{g}_{k+1} \\
&= \dots \\
&= W^{k+1}\mathbf{y}_0 - W^k\mathbf{g}_0 + \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= W^k(W-I)\mathbf{g}_0 + \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1} \\
&= \sum_{j=0}^{k} W^j(W-I)\mathbf{g}_{k-j} + \mathbf{g}_{k+1},
\end{aligned}$$

$$\begin{aligned}
\mathbf{y}_k - \mathbf{1}\bar{y}_k &= \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1}\frac{1}{n}\mathbf{1}\mathbf{1}^T W^j(W-I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1} W^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \sum_{j=0}^{k-1}\frac{1}{n}\mathbf{1}\mathbf{1}^T(W-I)\mathbf{g}_{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1}(W^j - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{g}_k \\
&= \sum_{j=0}^{k-1}(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j(W-I)\mathbf{g}_{k-1-j} + \mathbf{g}_k - \mathbf{1}\bar{g}_k,
\end{aligned}$$

where the last equality can be proven by recursion and the fact that the matrix W is doubly stochastic by Assumption 1.1:

$(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^{j+1} = (W^j - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = W^{j+1} - \frac{1}{n}W^j\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T W + \frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1} - \frac{2}{n}\mathbf{1}\mathbf{1}^T + \frac{1}{n}\mathbf{1}\mathbf{1}^T = W^{j+1} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Thus,

$$\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \sum_{j=0}^{k-1}\|(W - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^j(W-I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\| \\
&\leq \sum_{j=0}^{k-1}\rho_w^j\|(W-I)\mathbf{g}_{k-1-j}\| + \|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|.
\end{aligned}$$

From Lemma F.1, we have $\|\mathbf{g}_k\|^2 < \infty$ almost surely.

$$\begin{aligned}
\|\mathbf{g}_k - \mathbf{1}\bar{g}_k\|^2 &= \sum_{i=1}^{n} \|g_{i,k} - \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\|^2 \\
&= \sum_{i=1}^{n} \left(\|g_{i,k}\|^2 - 2\langle g_{i,k}, \frac{1}{n}\sum_{j=1}^{n} g_{j,k}\rangle + \|\bar{g}_k\|^2 \right) \\
&= \|\mathbf{g}_k\|^2 - 2n\|\bar{g}_k\|^2 + n\|\bar{g}_k\|^2 \\
&= \|\mathbf{g}_k\|^2 - n\|\bar{g}_k\|^2 \\
&\leq \|\mathbf{g}_k\|^2 \\
&\leq M'^2 < \infty.
\end{aligned}$$

Inserting in the previous inequality, we get

$$\begin{aligned}
\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\| &\leq \frac{M'}{1-\rho_w}\|(W-I)\| + M' \\
&= G < \infty,
\end{aligned} \tag{54}$$

where we have a geometric sum as $\rho_w < 1$.

Substituting the upper bound in (53),

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 &\leq \big(\frac{1+\rho_w^2}{2}\big)^{k+1}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \alpha^2 \frac{2\rho_w^2 G^2}{1-\rho_w^2}\sum_{j=0}^{k}\big(\frac{1+\rho_w^2}{2}\big)^{j+1} \\
&\overset{(a)}{\leq} \big(\frac{1+\rho_w^2}{2}\big)^{k+1}\|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2 + \alpha^2 \frac{2\rho_w^2(1+\rho_w^2)G^2}{(1-\rho_w^2)^2}
\end{aligned} \tag{55}$$

where (a) is due to the geometric sum with $\frac{1+\rho_w^2}{2} < 1$.

We conclude that $\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$ converges linearly almost surely.

2. **Proving $D_k = \mathbb{E}[\|\bar{x}_k - x^*\|^2]$ converges linearly**

We start by expressing the expected divergence in terms of its previous iteration.

$$\begin{aligned}
D_{k+1} =& \mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2] \\
=& \mathbb{E}[\|\bar{x}_k - \alpha\bar{g}_k - x^*\|^2] \\
=& D_k + \alpha^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha\mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k\rangle] \\
\overset{(a)}{=}& D_k + \alpha^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha_2\alpha\gamma\mathbb{E}[\langle \bar{x}_k - x^*, h(\mathbf{x}_k) + \bar{b}_k\rangle] \\
=& D_k + \alpha^2\mathbb{E}[\|\bar{g}_k\|^2] - 2\alpha_2\alpha\gamma\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k)\rangle] + 2\alpha_2\alpha\gamma\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle] \\
&- 2\alpha_2\alpha\gamma\mathbb{E}[\langle \bar{x}_k - x^*, \bar{b}_k\rangle]
\end{aligned} \tag{56}$$

where (a) is due to both $\mathbb{E}[e_k|\mathcal{H}_k] = 0$ and (50):

$$\begin{aligned}
\mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k\rangle] &= \mathbb{E}[\langle \bar{x}_k - x^*, \bar{g}_k - \mathbb{E}[\bar{g}_k|\mathcal{H}_k] + \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= \mathbb{E}[\langle \bar{x}_k - x^*, e_k\rangle] + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= \mathbb{E}_{\mathcal{H}_k}[\mathbb{E}[\langle \bar{x}_k - x^*, e_k\rangle|\mathcal{H}_k]] + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle] \\
&= 0 + \mathbb{E}[\langle \bar{x}_k - x^*, \mathbb{E}[\bar{g}_k|\mathcal{H}_k]\rangle].
\end{aligned}$$

From Lemma F.1 and (50), we have $\mathbb{E}[\|\bar{g}_k\|^2] < \bar{M}$ almost surely, with $\bar{M} = \frac{1}{n}M$ a bounded constant.

By the strong convexity in Assumption 2.6, we have

$$\begin{aligned}
-2\alpha_2\alpha\gamma\mathbb{E}[\langle \bar{x}_k - x^*, \mathcal{F}(\bar{x}_k)\rangle] &\leq -2\lambda\alpha_2\alpha\gamma\mathbb{E}[\|\bar{x}_k - x^*\|^2] \\
&= -2\lambda\alpha_2\alpha\gamma D_k.
\end{aligned} \tag{57}$$

Next, from Lemma 1.6, we have

$$\begin{aligned}2\alpha_2\alpha\gamma\langle\bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle &\leq 2\alpha_2\alpha\gamma\frac{L}{\sqrt{n}}\|\bar{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\| \\ &\overset{(a)}{\leq} \frac{\lambda\alpha_2\alpha\gamma}{2}\|\bar{x}_k - x^*\|^2 + 2\alpha_2\alpha\gamma\frac{L^2}{\lambda n}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2,\end{aligned}$$

where (a) is due to $2\sqrt{\epsilon} \times \frac{1}{\sqrt{\epsilon}}\langle a, b\rangle = 2\langle\sqrt{\epsilon}a, \frac{1}{\sqrt{\epsilon}}b\rangle \leq \epsilon\|a\|^2 + \frac{1}{\epsilon}\|b\|^2$. In (55), we let $R = \|\mathbf{x}_0 - \mathbf{1}\bar{x}_0\|^2$, $\breve{G} = \frac{2\rho_w^2(1+\rho_w^2)G^2}{(1-\rho_w^2)^2}$,

$$\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \leq \big(\frac{1+\rho_w^2}{2}\big)^k R + \alpha^2\breve{G}. \tag{58}$$

Hence,

$$2\alpha_2\alpha\gamma\mathbb{E}[\langle\bar{x}_k - x^*, \mathcal{F}(\bar{x}_k) - h(\mathbf{x}_k)\rangle] \leq \frac{\lambda\alpha_2\alpha\gamma}{2}D_k + 2\alpha_2\alpha\gamma\frac{L^2}{\lambda n}\Big[\big(\frac{1+\rho_w^2}{2}\big)^k R + \alpha^2\breve{G}\Big]. \tag{59}$$

From (51),

$$\begin{aligned}-2\alpha_2\alpha\gamma\mathbb{E}[\langle\bar{x}_k - x^*, \bar{b}_k\rangle] &\leq \frac{\lambda\alpha_2\alpha\gamma}{2}D_k + \frac{2\alpha_2\alpha\gamma}{\lambda}\mathbb{E}[\|\bar{b}_k\|^2] \\ &\leq \frac{\lambda\alpha_2\alpha\gamma}{2}D_k + \alpha\gamma^3\frac{\alpha_1^2\alpha_3^6}{2\alpha_2\lambda}\end{aligned} \tag{60}$$

Finally, by combining (56), (57), (59), and (60) and setting again $A = \lambda\alpha_2$, $B = \frac{2\alpha_2 L^2}{\lambda n}$, and $C = \frac{\alpha_1^2\alpha_3^6}{2\alpha_2\lambda}$, we get

$$D_{k+1} \leq (1 - A\alpha\gamma)D_k + B\alpha\gamma\Big[\big(\frac{1+\rho_w^2}{2}\big)^k R + \alpha^2\breve{G}\Big] + \bar{M}\alpha^2 + C\alpha\gamma^3 \tag{61}$$

Let $\varrho_1 = 1 - A\alpha\gamma$ and $\varrho_2 = \big(\frac{1+\rho_w^2}{2}\big)$. Then, assuming $\alpha\gamma < \frac{1}{A}$ and taking the telescoping sum

$$\begin{aligned}D_{k+1} \leq& \varrho_1^{k+1}D_0 + \alpha\gamma BR\sum_{i=0}^{k}\varrho_1^i\varrho_2^{k-i} + (\alpha^3\gamma B\breve{G} + \alpha^2\bar{M} + \alpha\gamma^3 C)\sum_{i=0}^{k}\varrho_1^i \\ =& \varrho_1^{k+1}D_0 + \alpha\gamma BR\sum_{i=0}^{k}\varrho_1^i\varrho_2^{k-i} + (\alpha^3\gamma B\breve{G} + \alpha^2\bar{M} + \alpha\gamma^3 C)\Big(\frac{1-\varrho_1^{k+1}}{1-\varrho_1}\Big) \\ =& \varrho_1^{k+1}D_0 + \alpha\gamma BR\sum_{i=0}^{k}\varrho_1^i\varrho_2^{k-i} + \Big(\alpha^2\frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2\frac{C}{A}\Big)(1-\varrho_1^{k+1}) \\ \leq& \varrho_1^{k+1}D_0 + \alpha\gamma BR\sum_{i=0}^{k}\varrho_1^i\varrho_2^{k-i} + \alpha^2\frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2\frac{C}{A}\end{aligned} \tag{62}$$

where in the last equality, we further imposed the step sizes to satisfy $\alpha < \gamma$.

In what follows, we discuss the summation in the second term of the inequality to avoid setting loose bounds. We know that this summation can be written as follows,

$$\sum_{i=0}^{k}\varrho_1^i\varrho_2^{k-i} = \sum_{i=0}^{k}\varrho_1^{k-i}\varrho_2^i \tag{63}$$

Thus, without imposing further assumptions on the step sizes, we consider the following function the two cases:

- When $\varrho_1 \leq \varrho_2$, we use the left hand side of the previous equality

$$\begin{aligned} D_{k+1} \leq & \varrho_1^{k+1} D_0 + \alpha\gamma BR\varrho_2^k \sum_{i=0}^{k} \varrho_1^i \varrho_2^{-i} + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \\ \leq & \varrho_1^{k+1} D_0 + \alpha\gamma BR\varrho_2^k \frac{1}{1-\frac{\varrho_1}{\varrho_2}} + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \\ = & \varrho_1^{k+1} D_0 + \varrho_2^{k+1} \frac{2\alpha\gamma BR}{2A\alpha\gamma + \rho_w^2 - 1} + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \end{aligned} \tag{64}$$

  Then, for arbitrary small step sizes satisfying $\alpha\gamma < \frac{1}{A}$ and $\alpha < \gamma$, D_k converges with the linear rate of $O\big(\varrho_2^k\big)$.

- When $\varrho_1 > \varrho_2$, we use the right hand side

$$\begin{aligned} D_{k+1} \leq & \varrho_1^{k+1} D_0 + \alpha\gamma BR\varrho_1^k \sum_{i=0}^{k} \varrho^{-i} \varrho_2^i + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \\ \leq & \varrho_1^{k+1} D_0 + \alpha\gamma BR\varrho_1^k \frac{1}{1-\frac{\varrho_2}{\varrho_1}} + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \\ = & \varrho_1^{k+1} \Big(D_0 + \frac{2\alpha\gamma BR}{1 - 2A\alpha\gamma - \rho_w^2} \Big) + \alpha^2 \frac{B\breve{G}}{A} + \frac{\alpha}{\gamma}\frac{\bar{M}}{A} + \gamma^2 \frac{C}{A} \end{aligned} \tag{65}$$

  Then, for arbitrary small step sizes satisfying $\alpha\gamma < \frac{1}{A}$ and $\alpha < \gamma$, D_k converges with the linear rate of $O\big(\varrho_1^k\big)$.

G Additional Numerical Examples

We adjust the step sizes of 1P-ZOFL to $\alpha = 0.05$ and $\gamma = 0.25$ when they are fixed and $\alpha_k = 0.08(k+1)^{-0.75}$ and $\gamma_k = 0.7(k+1)^{-0.25}$ for classifying images with the labels 2 and 3 in Figures 8-11.

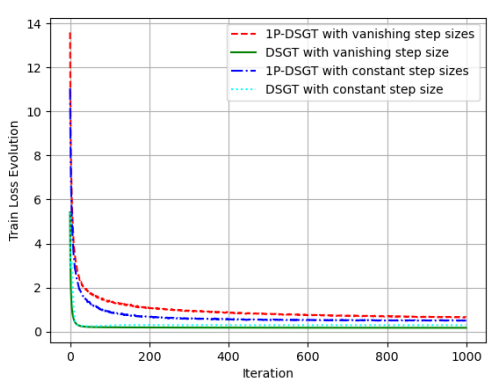

Figure 8: Expected loss function evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 2 and 3.

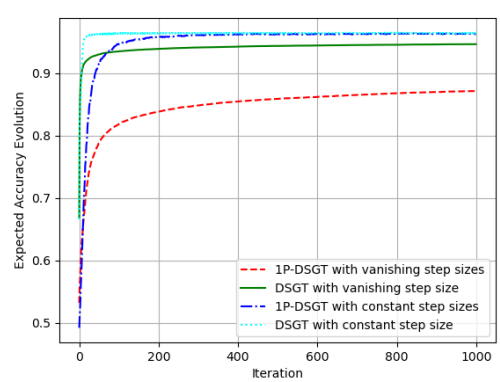

Figure 9: Expected test accuracy evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 2 and 3.

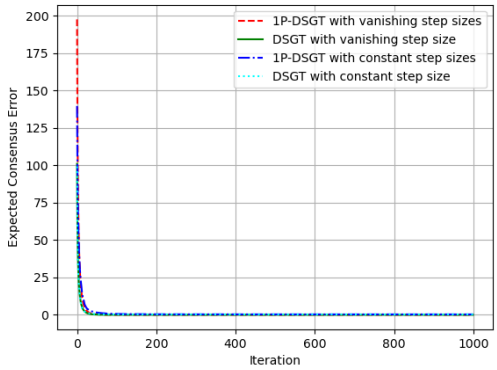

Figure 10: Expected consensus error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 2 and 3.

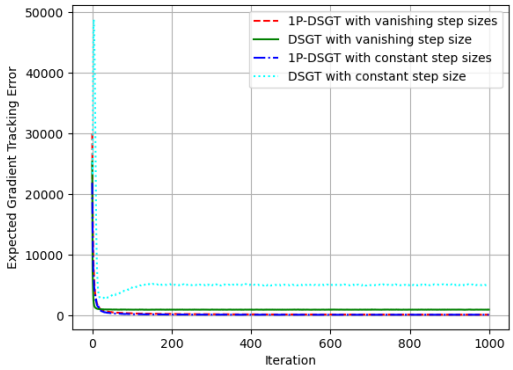

Figure 11: Expected gradient tracking error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 2 and 3.

In Figures 12-15 depicting the classification of images with the labels 3 and 4, we adjust the step sizes of 1P-ZOFL to $\alpha = 0.02$ and $\gamma = 0.2$ when they are fixed and $\alpha_k = 0.08(k+1)^{-0.75}$ and $\gamma_k = 0.7(k+1)^{-0.25}$.

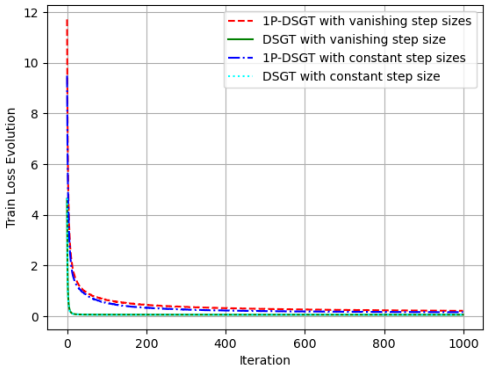

Figure 12: Expected loss function evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 3 and 4.

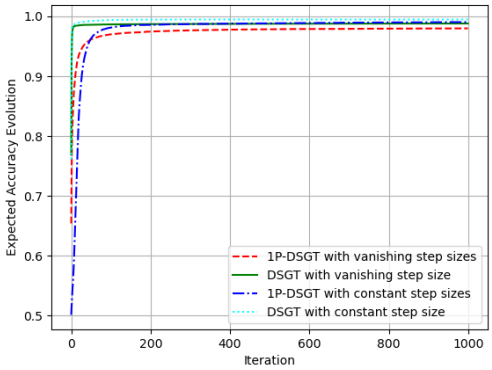

Figure 13: Expected test accuracy evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 3 and 4.

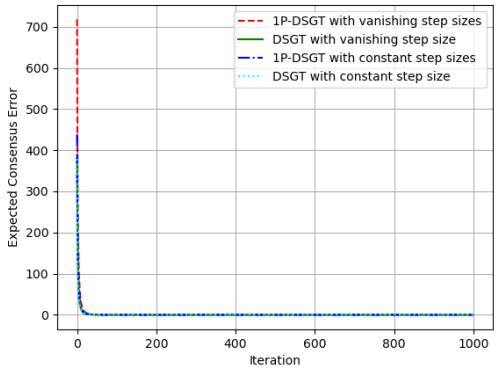

Figure 14: Expected consensus error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 3 and 4.

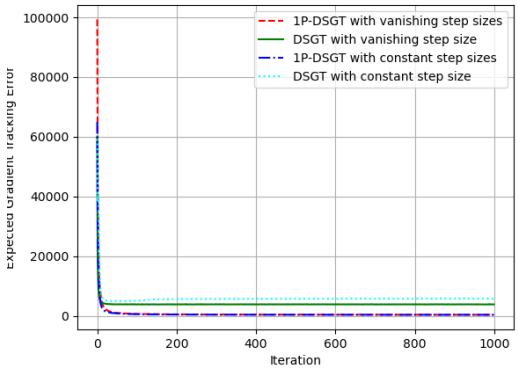

Figure 15: Expected gradient tracking error evolution of the algorithms 1P-DSGT vs. DSGT considering vanishing vs. constant step sizes classifying images with labels 3 and 4.